ECONOMETRICS

Bruce E. Hansen

University of Wisconsin

Department of Economics

This Revision: January 2018
Comments Welcome

Contents

Preface

This book is intended to serve as the textbook a first-year graduate course in econometrics.

Students are assumed to have an understanding of multivariate calculus, probability theory, linear algebra, and mathematical statistics. A prior course in undergraduate econometrics would be helpful, but not required. Two excellent undergraduate textbooks are Wooldridge (2015) and Stock and Watson (2014).

For reference, some of the basic tools of matrix algebra and probability inequalites are reviewed in the Appendix.

For students wishing to deepen their knowledge of matrix algebra in relation to their study of econometrics, I recommend *Matrix Algebra* by Abadir and Magnus (2005).

An excellent introduction to probability and statistics is *Statistical Inference* by Casella and Berger (2002). For those wanting a deeper foundation in probability, I recommend Ash (1972) or Billingsley (1995). For more advanced statistical theory, I recommend Lehmann and Casella (1998), van der Vaart (1998), Shao (2003), and Lehmann and Romano (2005).

For further study in econometrics beyond this text, I recommend Davidson (1994) for asymptotic theory, Hamilton (1994) and Kilian and Lütkepohl (2017) for time-series methods, Wooldridge (2010) for panel data and discrete response models, and Li and Racine (2007) for nonparametrics and semiparametric econometrics. Beyond these texts, the *Handbook of Econometrics* series provides advanced summaries of contemporary econometric methods and theory.

The end-of-chapter exercises are important parts of the text and are meant to help teach students of econometrics. Answers are not provided, and this is intentional.

I would like to thank Ying-Ying Lee and Wooyoung Kim for providing research assistance in preparing some of the empirical examples presented in the text.

This is a manuscript in progress. Chapters 1-11 are mostly complete. Chapters 12-18 are incomplete.

Chapter 1

Introduction

1.1 What is Econometrics?

The term "econometrics" is believed to have been crafted by Ragnar Frisch (1895-1973) of Norway, one of the three principal founders of the Econometric Society, first editor of the journal *Econometrica*, and co-winner of the first Nobel Memorial Prize in Economic Sciences in 1969. It is therefore fitting that we turn to Frisch's own words in the introduction to the first issue of *Econometrica* to describe the discipline.

> A word of explanation regarding the term econometrics may be in order. Its definition is implied in the statement of the scope of the [Econometric] Society, in Section I of the Constitution, which reads: "The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics.... Its main object shall be to promote studies that aim at a unification of the theoretical-quantitative and the empirical-quantitative approach to economic problems...."
>
> But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a defininitely quantitative character. Nor should econometrics be taken as synonomous with the application of mathematics to economics. Experience has shown that each of these three view-points, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.
>
> Ragnar Frisch, *Econometrica*, (1933), 1, pp. 1-2.

This definition remains valid today, although some terms have evolved somewhat in their usage. Today, we would say that econometrics is the unified study of economic models, mathematical statistics, and economic data.

Within the field of econometrics there are sub-divisions and specializations. **Econometric theory** concerns the development of tools and methods, and the study of the properties of econometric methods. **Applied econometrics** is a term describing the development of quantitative economic models and the application of econometric methods to these models using economic data.

1.2 The Probability Approach to Econometrics

The unifying methodology of modern econometrics was articulated by Trygve Haavelmo (1911-1999) of Norway, winner of the 1989 Nobel Memorial Prize in Economic Sciences, in his seminal

paper "The probability approach in econometrics" (1944). Haavelmo argued that quantitative economic models must necessarily be *probability models* (by which today we would mean *stochastic*). Deterministic models are blatently inconsistent with observed economic quantities, and it is incoherent to apply deterministic models to non-deterministic data. Economic models should be explicitly designed to incorporate randomness; stochastic errors should not be simply added to deterministic models to make them random. Once we acknowledge that an economic model is a probability model, it follows naturally that an appropriate tool way to quantify, estimate, and conduct inferences about the economy is through the powerful theory of mathematical statistics. The appropriate method for a quantitative economic analysis follows from the probabilistic construction of the economic model.

Haavelmo's probability approach was quickly embraced by the economics profession. Today no quantitative work in economics shuns its fundamental vision.

While all economists embrace the probability approach, there has been some evolution in its implementation.

The **structural approach** is the closest to Haavelmo's original idea. A probabilistic economic model is specified, and the quantitative analysis performed under the assumption that the economic model is correctly specified. Researchers often describe this as "taking their model seriously." The structural approach typically leads to likelihood-based analysis, including maximum likelihood and Bayesian estimation.

A criticism of the structural approach is that it is misleading to treat an economic model as correctly specified. Rather, it is more accurate to view a model as a useful abstraction or approximation. In this case, how should we interpret structural econometric analysis? The **quasi-structural approach** to inference views a structural economic model as an approximation rather than the truth. This theory has led to the concepts of the pseudo-true value (the parameter value defined by the estimation problem), the quasi-likelihood function, quasi-MLE, and quasi-likelihood inference.

Closely related is the **semiparametric approach**. A probabilistic economic model is partially specified but some features are left unspecified. This approach typically leads to estimation methods such as least-squares and the Generalized Method of Moments. The semiparametric approach dominates contemporary econometrics, and is the main focus of this textbook.

Another branch of quantitative structural economics is the **calibration approach.** Similar to the quasi-structural approach, the calibration approach interprets structural models as approximations and hence inherently false. The difference is that the calibrationist literature rejects mathematical statistics (deeming classical theory as inappropriate for approximate models) and instead selects parameters by matching model and data moments using non-statistical *ad hoc*[1] methods.

1.3 Econometric Terms and Notation

In a typical application, an econometrician has a set of repeated measurements on a set of variables. For example, in a labor application the variables could include weekly earnings, educational attainment, age, and other descriptive characteristics. We call this information the **data, dataset**, or **sample**.

We use the term **observations** to refer to the distinct repeated measurements on the variables. An individual observation often corresponds to a specific economic unit, such as a person, household, corporation, firm, organization, country, state, city or other geographical region. An individual observation could also be a measurement at a point in time, such as quarterly GDP or a daily interest rate.

[1] *Ad hoc* means "for this purpose" – a method designed for a specific problem – and not based on a generalizable principle.

Economists typically denote variables by the italicized roman characters y, x, and/or z. The convention in econometrics is to use the character y to denote the variable to be explained, while the characters x and z are used to denote the conditioning (explaining) variables.

Following mathematical convention, real numbers (elements of the real line $\mathbb{R}$, also called **scalars**) are written using lower case italics such as y, and vectors (elements of $\mathbb{R}^k$) by lower case bold italics such as $\boldsymbol{x}$, e.g.

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}.$$

Upper case bold italics such as $\boldsymbol{X}$ are used for matrices.

We denote the number of observations by the natural number n, and subscript the variables by the index i to denote the individual observation, e.g. y_i, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. In some contexts we use indices other than i, such as in time-series applications where the index t is common and T is used to denote the number of observations. In panel studies we typically use the double index it to refer to individual i at a time period t.

The i^{th} **observation** is the set $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. The **sample** is the set $\{(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i) : i = 1, ..., n\}$.

It is proper mathematical practice to use upper case X for random variables and lower case x for realizations or specific values. Since we use upper case to denote matrices, the distinction between random variables and their realizations is not rigorously followed in econometric notation. Thus the notation y_i will in some places refer to a random variable, and in other places a specific realization. This is undesirable but there is little to be done about it without terrifically complicating the notation. Hopefully there will be no confusion as the use should be evident from the context.

We typically use Greek letters such as β, θ and σ^2 to denote unknown parameters of an econometric model, and will use boldface, e.g. $\boldsymbol{\beta}$ or $\boldsymbol{\theta}$, when these are vector-valued. Estimates are typically denoted by putting a hat "^", tilde "~" or bar "-" over the corresponding letter, e.g. $\widehat{\beta}$ and $\widetilde{\beta}$ are estimates of β.

The covariance matrix of an econometric estimator will typically be written using the capital boldface $\boldsymbol{V}$, often with a subscript to denote the estimator, e.g. $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \text{var}\left(\widehat{\boldsymbol{\beta}}\right)$ as the covariance matrix for $\widehat{\boldsymbol{\beta}}$. Hopefully without causing confusion, we will use the notation $\boldsymbol{V}_{\boldsymbol{\beta}} = \text{avar}(\widehat{\boldsymbol{\beta}})$ to denote the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ (the variance of the asymptotic distribution). Estimates will be denoted by appending hats or tildes, e.g. $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{V}_{\boldsymbol{\beta}}$.

1.4 Observational Data

A common econometric question is to quantify the impact of one set of variables on another variable. For example, a concern in labor economics is the returns to schooling – the change in earnings induced by increasing a worker's education, holding other variables constant. Another issue of interest is the earnings gap between men and women.

Ideally, we would use **experimental** data to answer these questions. To measure the returns to schooling, an experiment might randomly divide children into groups, mandate different levels of education to the different groups, and then follow the children's wage path after they mature and enter the labor force. The differences between the groups would be direct measurements of the effects of different levels of education. However, experiments such as this would be widely

condemned as immoral! Consequently, in economics non-laboratory experimental data sets are typically narrow in scope.

Instead, most economic data is **observational**. To continue the above example, through data collection we can record the level of a person's education and their wage. With such data we can measure the joint distribution of these variables, and assess the joint dependence. But from observational data it is difficult to infer **causality**, as we are not able to manipulate one variable to see the direct effect on the other. For example, a person's level of education is (at least partially) determined by that person's choices. These factors are likely to be affected by their personal abilities and attitudes towards work. The fact that a person is highly educated suggests a high level of ability, which suggests a high relative wage. This is an alternative explanation for an observed positive correlation between educational levels and wages. High ability individuals do better in school, and therefore choose to attain higher levels of education, and their high ability is the fundamental reason for their high wages. The point is that multiple explanations are consistent with a positive correlation between schooling levels and education. Knowledge of the joint distribution alone may not be able to distinguish between these explanations.

Most economic data sets are observational, not experimental. This means that all variables must be treated as random and possibly jointly determined.

This discussion means that it is difficult to infer causality from observational data alone. Causal inference requires identification, and this is based on strong assumptions. We will discuss these issues on occasion throughout the text.

1.5 Standard Data Structures

There are five major types of economic data sets: cross-sectional, time-series, panel, clustered, and spatial. They are distinguished by the dependence structure across observations.

Cross-sectional data sets have one observation per individual. Surveys and administrative records are a typical source for cross-sectional data. In typical applications, the individuals surveyed are persons, households, firms or other economic agents. In many contemporary econometric cross-section studies the sample size n is quite large. It is conventional to assume that cross-sectional observations are mutually independent. Most of this text is devoted to the study of cross-section data.

Time-series data are indexed by time. Typical examples include macroeconomic aggregates, prices and interest rates. This type of data is characterized by serial dependence. Most aggregate economic data is only available at a low frequency (annual, quarterly or perhaps monthly) so the sample size is typically much smaller than in cross-section studies. An exception is financial data where data are available at a high frequency (weekly, daily, hourly, or by transaction) so sample sizes can be quite large.

Panel data combines elements of cross-section and time-series. These data sets consist of a set of individuals (typically persons, households, or corporations) measured repeatedly over time. The common modeling assumption is that the individuals are mutually independent of one another, but a given individual's observations are mutually dependent. In some panel data contexts, the number of time series observations T per individual is small while the number of individuals n is large. In other panel data contexts (for example when countries or states are taken as the unit of measurement) the number of individuals n can be small while the number of time series observations T can be moderately large. An important issue in econometric panel data is the treatment of error components.

Clustered samples are increasing popular in applied economics, and is related to panel data. In clustered sampling, the observations are grouped into "clusters" which are treated as mutually independent, yet allowed to be dependent within the cluster. The major difference with panel data is that clustered sampling typically does not explicitly model error component structures, nor the dependence within clusters, but rather is concerned with inference which is robust to arbitrary forms of within-cluster correlation.

Spatial dependence is another model of interdependence. The observations are treated as mutually dependent according to a spatial measure (for example, geographic proximity). Unlike clustering, spatial models allow all observations to be mutually dependent, and typically rely on explicit modeling of the dependence relationships. Spatial dependence can also be viewed as a generalization of time series dependence.

Data Structures

- Cross-section
- Time-series
- Panel
- Clustered
- Spatial

As we mentioned above, most of this text will be devoted to cross-sectional data under the assumption of mutually independent observations. By mutual independence we mean that the i^{th} observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ is independent of the j^{th} observation $(y_j, \boldsymbol{x}_j, \boldsymbol{z}_j)$ for $i \neq j$. (Sometimes the label "independent" is misconstrued. It is a statement about the relationship between observations i and j, not a statement about the relationship between y_i and $\boldsymbol{x}_i$ and/or $\boldsymbol{z}_i$.) In this case we say that the data are **independently distributed**.

Furthermore, if the data is randomly gathered, it is reasonable to model each observation as a draw from the same probability distribution. In this case we say that the data are **identically distributed**. If the observations are mutually independent and identically distributed, we say that the observations are **independent and identically distributed**, **iid**, or a **random sample**. For most of this text we will assume that our observations come from a random sample.

Definition 1.5.1 *The observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ are a **sample** from the distribution F if they are identically distributed across $i = 1, ..., n$ with joint distribution F.*

Definition 1.5.2 *The observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ are a **random sample** if they are mutually independent and identically distributed (**iid**) across $i = 1, ..., n$.*

In the random sampling framework, we think of an individual observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ as a realization from a joint probability distribution $F(y, \boldsymbol{x}, \boldsymbol{z})$ which we can call the **population**. This "population" is infinitely large. This abstraction can be a source of confusion as it does not correspond to a physical population in the real world. It is an abstraction since the distribution F is unknown, and the goal of statistical inference is to learn about features of F from the sample. The *assumption* of random sampling provides the mathematical foundation for treating economic statistics with the tools of mathematical statistics.

The random sampling framework was a major intellectual breakthrough of the late 19th century, allowing the application of mathematical statistics to the social sciences. Before this conceptual development, methods from mathematical statistics had not been applied to economic data as the latter was viewed as non-random. The random sampling framework enabled economic samples to be treated as random, a necessary precondition for the application of statistical methods.

1.6 Sources for Economic Data

Fortunately for economists, the internet provides a convenient forum for dissemination of economic data. Many large-scale economic datasets are available without charge from governmental agencies. An excellent starting point is the Resources for Economists Data Links, available at `rfe.org`. From this site you can find almost every publically available economic data set. Some specific data sources of interest include

- Bureau of Labor Statistics
- US Census
- Current Population Survey
- Survey of Income and Program Participation
- Panel Study of Income Dynamics
- Federal Reserve System (Board of Governors and regional banks)
- National Bureau of Economic Research
- U.S. Bureau of Economic Analysis
- CompuStat
- International Financial Statistics

Another good source of data is from authors of published empirical studies. Most journals in economics require authors of published papers to make their datasets generally available. For example, in its instructions for submission, *Econometrica* states:

> *Econometrica* has the policy that all empirical, experimental and simulation results must be replicable. Therefore, authors of accepted papers must submit data sets, programs, and information on empirical analysis, experiments and simulations that are needed for replication and some limited sensitivity analysis.

The *American Economic Review* states:

> All data used in analysis must be made available to any researcher for purposes of replication.

The *Journal of Political Economy* states:

> It is the policy of the *Journal of Political Economy* to publish papers only if the data used in the analysis are clearly and precisely documented and are readily available to any researcher for purposes of replication.

If you are interested in using the data from a published paper, first check the journal's website, as many journals archive data and replication programs online. Second, check the website(s) of the paper's author(s). Most academic economists maintain webpages, and some make available replication files complete with data and programs. If these investigations fail, email the author(s), politely requesting the data. You may need to be persistent.

As a matter of professional etiquette, all authors absolutely have the obligation to make their data and programs available. Unfortunately, many fail to do so, and typically for poor reasons. The irony of the situation is that it is typically in the best interests of a scholar to make as much of their work (including all data and programs) freely available, as this only increases the likelihood of their work being cited and having an impact.

Keep this in mind as you start your own empirical project. Remember that as part of your end product, you will need (and want) to provide all data and programs to the community of scholars. The greatest form of flattery is to learn that another scholar has read your paper, wants to extend your work, or wants to use your empirical methods. In addition, public openness provides a healthy incentive for transparency and integrity in empirical analysis.

1.7 Econometric Software

Economists use a variety of econometric, statistical, and programming software.

Stata (www.stata.com) is a powerful statistical program with a broad set of pre-programmed econometric and statistical tools. It is quite popular among economists, and is continuously being updated with new methods. It is an excellent package for most econometric analysis, but is limited when you want to use new or less-common econometric methods which have not yet been programed.

R (www.r-project.org), GAUSS (www.aptech.com), MATLAB (www.mathworks.com), and Ox-Metrics (www.oxmetrics.net) are high-level matrix programming languages with a wide variety of built-in statistical functions. Many econometric methods have been programed in these languages and are available on the web. The advantage of these packages is that you are in complete control of your analysis, and it is easier to program new methods than in Stata. Some disadvantages are that you have to do much of the programming yourself, programming complicated procedures takes significant time, and programming errors are hard to prevent and difficult to detect and eliminate. Of these languages, GAUSS used to be quite popular among econometricians, but currently MATLAB is more popular. A smaller but growing group of econometricians are enthusiastic fans of R, which of these languages is uniquely open-source, user-contributed, and best of all, completely free!

For highly-intensive computational tasks, some economists write their programs in a standard programming language such as Fortran or C. This can lead to major gains in computational speed, at the cost of increased time in programming and debugging.

As these different packages have distinct advantages, many empirical economists end up using more than one package. As a student of econometrics, you will learn at least one of these packages, and probably more than one.

1.8 Data Files for Textbook

On the textbook webpage http://www.ssc.wisc.edu/~bhansen/econometrics/ there are posted a number of files containing data sets which are used in this textbook both for illustration and for end-of-chapter empirical exercises. For each data sets there are four files: (1) Description (pdf format); (2) Excel data file; (3) Text data file; (4) Stata data file. The three data files are identical

in content, the observations and variables are listed in the same order in each, all have variable labels.

For example, the text makes frequent reference to a wage data set extracted from the Current Population Survey. This data set is named `cps09mar`, and is represented by the files `cps09mar_description.pdf`, `cps09mar.xlsx`, `cps09mar.txt`, and `cps09mar.dta`.

The data sets currently included are

- `cps09mar`
  - household survey data extracted from the March 2009 Current Population Survey
- `DDK2011`
  - Data file from Duflo, Dupas and Kremer (2011)
- `invest`
  - Data file from B.E. Hansen (1999), extracted from Hall and Hall (1993)
- `Nerlove1963`
  - Data file from Nerlov (1963)
- `MRW1992`
  - Data file from Mankiw, Romer and Weil (1992)
- `Card1995`
  - Data file from Card (1995)
- `AJR2001`
  - Data file from Acemoglu, Johnson and Robinson (2001)
- `AK1991`
  - Data file from Angrist and Krueger (1991)
- `hprice1`
  - Housing price data. The only files posted are `hprice1.txt` and `hprice1.pdf` which are the data in text format and description, respectively

1.9 Reading the Manuscript

I have endeavored to use a unified notation and nomenclature. The development of the material is cumulative, with later chapters building on the earlier ones. Nevertheless, every attempt has been made to make each chapter self-contained, so readers can pick and choose topics according to their interests.

To fully understand econometric methods, it is necessary to have a mathematical understanding of its mechanics, and this includes the mathematical proofs of the main results. Consequently, this text is self-contained, with nearly all results proved with full mathematical rigor. The mathematical development and proofs aim at brevity and conciseness (sometimes described as mathematical

elegance), but also at pedagogy. To understand a mathematical proof, it is not sufficient to simply *read* the proof, you need to follow it, and re-create it for yourself.

Nevertheless, many readers will not be interested in each mathematical detail, explanation, or proof. This is okay. To use a method it may not be necessary to understand the mathematical details. Accordingly I have placed the more technical mathematical proofs and details in chapter appendices. These appendices and other technical sections are marked with an asterisk (*). These sections can be skipped without any loss in exposition.

1.10 Common Symbols

y	scalar
$\boldsymbol{x}$	vector
$\boldsymbol{X}$	matrix
$\mathbb{R}$	real line
$\mathbb{R}^k$	Euclidean k space
$\mathbb{E}(y)$	mathematical expectation
$\text{var}(y)$	variance
$\text{cov}(x, y)$	covariance
$\text{var}(\boldsymbol{x})$	covariance matrix
$\text{corr}(x, y)$	correlation
Pr	probability
$\longrightarrow$	limit
$\xrightarrow{p}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
$\text{plim}_{n\to\infty}$	probability limit
$\text{N}(0, 1)$	standard normal distribution
$\text{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
χ^2_k	chi-square distribution with k degrees of freedom
$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\text{tr}\,\boldsymbol{A}$	trace
$\boldsymbol{A}'$	matrix transpose
$\boldsymbol{A}^{-1}$	matrix inverse
$\boldsymbol{A} > 0$	positive definite
$\boldsymbol{A} \geq 0$	positive semi-definite
$\|\boldsymbol{a}\|$	Euclidean norm
$\|\boldsymbol{A}\|$	matrix (Frobinius or spectral) norm
$\approx$	approximate equality
$\stackrel{def}{=}$	definitional equality
$\sim$	is distributed as
log	natural logarithm

Chapter 2

Conditional Expectation and Projection

2.1 Introduction

The most commonly applied econometric tool is least-squares estimation, also known as **regression**. As we will see, least-squares is a tool to estimate an approximate conditional mean of one variable (the **dependent variable**) given another set of variables (the **regressors**, **conditioning variables**, or **covariates**).

In this chapter we abstract from estimation, and focus on the probabilistic foundation of the conditional expectation model and its projection approximation.

2.2 The Distribution of Wages

Suppose that we are interested in wage rates in the United States. Since wage rates vary across workers, we cannot describe wage rates by a single number. Instead, we can describe wages using a probability distribution. Formally, we view the wage of an individual worker as a random variable *wage* with the **probability distribution**

$$F(u) = \Pr(wage \leq u).$$

When we say that a person's wage is random we mean that we do not know their wage before it is measured, and we treat observed wage rates as realizations from the distribution F. Treating unobserved wages as random variables and observed wages as realizations is a powerful mathematical abstraction which allows us to use the tools of mathematical probability.

A useful thought experiment is to imagine dialing a telephone number selected at random, and then asking the person who responds to tell us their wage rate. (Assume for simplicity that all workers have equal access to telephones, and that the person who answers your call will respond honestly.) In this thought experiment, the wage of the person you have called is a single draw from the distribution F of wages in the population. By making many such phone calls we can learn the distribution F of the entire population.

When a distribution function F is differentiable we define the probability density function

$$f(u) = \frac{d}{du}F(u).$$

The density contains the same information as the distribution function, but the density is typically easier to visually interpret.

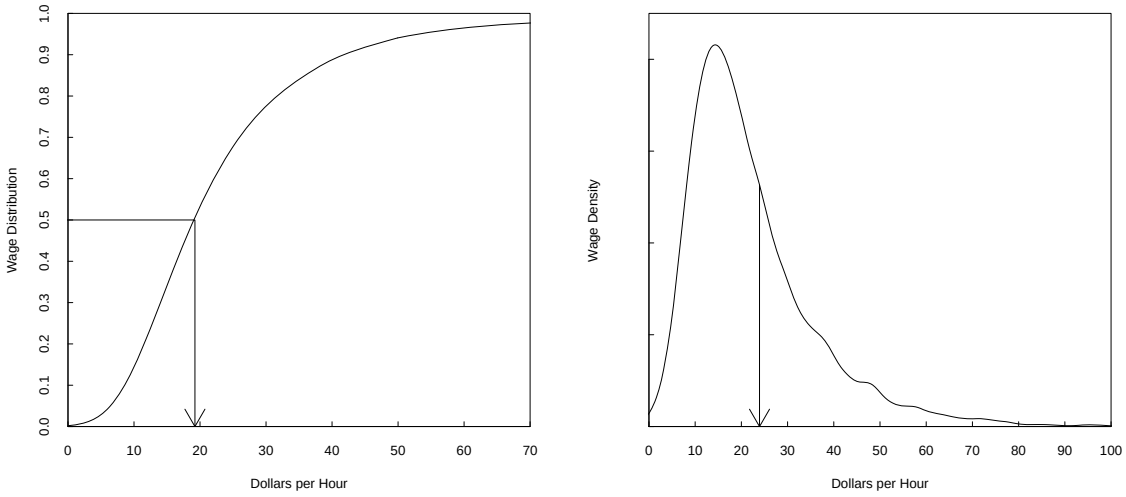


Figure 2.1: Wage Distribution and Density. All full-time U.S. workers

In Figure 2.1 we display estimates[1] of the probability distribution function (on the left) and density function (on the right) of U.S. wage rates in 2009. We see that the density is peaked around \$15, and most of the probability mass appears to lie between \$10 and \$40. These are ranges for typical wage rates in the U.S. population.

Important measures of central tendency are the median and the mean. The **median** m of a continuous[2] distribution F is the unique solution to

$$F(m) = \frac{1}{2}.$$

The median U.S. wage (\$19.23) is indicated in the left panel of Figure 2.1 by the arrow. The median is a robust[3] measure of central tendency, but it is tricky to use for many calculations as it is not a linear operator.

The **expectation** or **mean** of a random variable y with density f is

$$\mu = \mathbb{E}(y) = \int_{-\infty}^{\infty} uf(u)du.$$

Here we have used the common and convenient convention of using the single character y to denote a random variable, rather than the more cumbersome label *wage*. A general definition of the mean is presented in Section 2.30. The mean U.S. wage (\$23.90) is indicated in the right panel of Figure 2.1 by the arrow.

We sometimes use the notation $\mathbb{E}y$ instead of $\mathbb{E}(y)$ when the variable whose expectation is being taken is clear from the context. There is no distinction in meaning.

The mean is a convenient measure of central tendency because it is a linear operator and arises naturally in many economic models. A disadvantage of the mean is that it is not robust[4] especially in the presence of substantial skewness or thick tails, which are both features of the wage

[1] The distribution and density are estimated nonparametrically from the sample of 50,742 full-time non-military wage-earners reported in the March 2009 Current Population Survey. The wage rate is constructed as annual individual wage and salary earnings divided by hours worked.

[2] If F is not continuous the definition is $m = \inf\{u : F(u) \geq \frac{1}{2}\}$

[3] The median is not sensitive to pertubations in the tails of the distribution.

[4] The mean is sensitive to pertubations in the tails of the distribution.

distribution as can be seen easily in the right panel of Figure 2.1. Another way of viewing this is that 64% of workers earn less that the mean wage of $23.90, suggesting that it is incorrect to describe the mean as a "typical" wage rate.

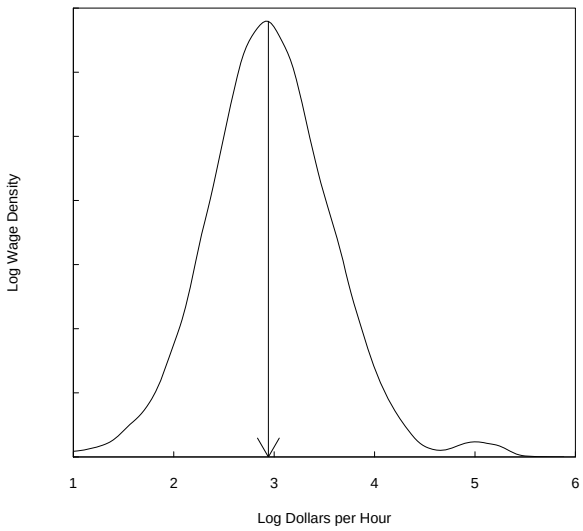


Figure 2.2: Log Wage Density

In this context it is useful to transform the data by taking the natural logarithm[5]. Figure 2.2 shows the density of log hourly wages $\log(wage)$ for the same population, with its mean 2.95 drawn in with the arrow. The density of log wages is much less skewed and fat-tailed than the density of the level of wages, so its mean

$$\mathbb{E}(\log(wage)) = 2.95$$

is a much better (more robust) measure[6] of central tendency of the distribution. For this reason, wage regressions typically use log wages as a dependent variable rather than the level of wages.

Another useful way to summarize the probability distribution $F(u)$ is in terms of its quantiles. For any $\alpha \in (0,1)$, the α^{th} quantile of the continuous[7] distribution F is the real number q_α which satisfies

$$F(q_\alpha) = \alpha.$$

The quantile function q_α, viewed as a function of α, is the inverse of the distribution function F. The most commonly used quantile is the median, that is, $q_{0.5} = m$. We sometimes refer to quantiles by the percentile representation of α, and in this case they are often called percentiles, e.g. the median is the 50^{th} percentile.

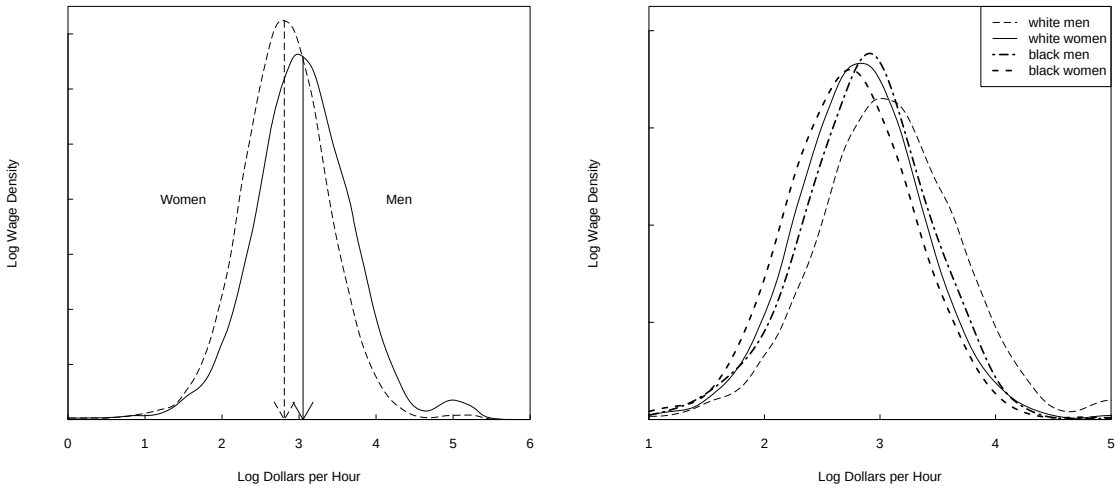
2.3 Conditional Expectation

We saw in Figure 2.2 the density of log wages. Is this distribution the same for all workers, or does the wage distribution vary across subpopulations? To answer this question, we can compare wage distributions for different groups – for example, men and women. The plot on the left in Figure 2.3 displays the densities of log wages for U.S. men and women with their means (3.05 and 2.81) indicated by the arrows. We can see that the two wage densities take similar shapes but the density for men is somewhat shifted to the right with a higher mean.

[5] Throughout the text, we will use $\log(y)$ or $\log y$ to denote the natural logarithm of y.

[6] More precisely, the geometric mean $\exp(\mathbb{E}(\log w)) = \$19.11$ is a robust measure of central tendency.

[7] If F is not continuous the definition is $q_\alpha = \inf\{u : F(u) \geq \alpha\}$



(a) Women and Men

(b) By Sex and Race

Figure 2.3: Log Wage Density by Sex and Race

The values 3.05 and 2.81 are the mean log wages in the subpopulations of men and women workers. They are called the **conditional means** (or **conditional expectations**) of log wages given sex. We can write their specific values as

$$\mathbb{E}\left(\log(wage) \mid sex = man\right) = 3.05 \tag{2.1}$$

$$\mathbb{E}\left(\log(wage) \mid sex = woman\right) = 2.81. \tag{2.2}$$

We call these means *conditional* as they are conditioning on a fixed value of the variable *sex*. While you might not think of a person's sex as a random variable, it is random from the viewpoint of econometric analysis. If you randomly select an individual, the sex of the individual is unknown and thus random. (In the population of U.S. workers, the probability that a worker is a woman happens to be 43%.) In observational data, it is most appropriate to view all measurements as random variables, and the means of subpopulations are then conditional means.

As the two densities in Figure 2.3 appear similar, a hasty inference might be that there is not a meaningful difference between the wage distributions of men and women. Before jumping to this conclusion let us examine the differences in the distributions of Figure 2.3 more carefully. As we mentioned above, the primary difference between the two densities appears to be their means. This difference equals

$$\begin{aligned}\mathbb{E}\left(\log(wage) \mid sex = man\right) - \mathbb{E}\left(\log(wage) \mid sex = woman\right) &= 3.05 - 2.81 \\ &= 0.24. \end{aligned} \tag{2.3}$$

A difference in expected log wages of 0.24 implies an average 24% difference between the wages of men and women, which is quite substantial. (For an explanation of logarithmic and percentage differences see Section 2.4.)

Consider further splitting the men and women subpopulations by race, dividing the population into whites, blacks, and other races. We display the log wage density functions of four of these groups on the right in Figure 2.3. Again we see that the primary difference between the four density functions is their central tendency.

	men	women
white	3.07	2.82
black	2.86	2.73
other	3.03	2.86

Table 2.1: Mean Log Wages by Sex and Race

Focusing on the means of these distributions, Table 2.1 reports the mean log wage for each of the six sub-populations.

The entries in Table 2.1 are the conditional means of $\log(wage)$ given *sex* and *race*. For example

$$\mathbb{E}\left(\log(wage) \mid sex = man,\ race = white\right) = 3.07$$

and

$$\mathbb{E}\left(\log(wage) \mid sex = woman,\ race = black\right) = 2.73.$$

One benefit of focusing on conditional means is that they reduce complicated distributions to a single summary measure, and thereby facilitate comparisons across groups. Because of this simplifying property, conditional means are the primary interest of regression analysis and are a major focus in econometrics.

Table 2.1 allows us to easily calculate average wage differences between groups. For example, we can see that the wage gap between men and women continues after disaggregation by race, as the average gap between white men and white women is 25%, and that between black men and black women is 13%. We also can see that there is a race gap, as the average wages of blacks are substantially less than the other race categories. In particular, the average wage gap between white men and black men is 21%, and that between white women and black women is 9%.

2.4 Log Differences*

A useful approximation for the natural logarithm for small x is

$$\log(1 + x) \approx x. \tag{2.4}$$

This can be derived from the infinite series expansion of $\log(1 + x)$:

$$\begin{aligned} \log(1 + x) &= x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots \\ &= x + O(x^2). \end{aligned}$$

The symbol $O(x^2)$ means that the remainder is bounded by Ax^2 as $x \to 0$ for some $A < \infty$. A plot of $\log(1 + x)$ and the linear approximation x is shown in Figure 2.4. We can see that $\log(1 + x)$ and the linear approximation x are very close for $|x| \leq 0.1$, and reasonably close for $|x| \leq 0.2$, but the difference increases with $|x|$.

Now, if y^* is $c\%$ greater than y, then

$$y^* = (1 + c/100)y.$$

Taking natural logarithms,

$$\log y^* = \log y + \log(1 + c/100)$$

or

$$\log y^* - \log y = \log(1 + c/100) \approx \frac{c}{100}$$

where the approximation is (2.4). This shows that 100 multiplied by the difference in logarithms is approximately the percentage difference between y and y^*, and this approximation is quite good for $|c| \leq 10$.

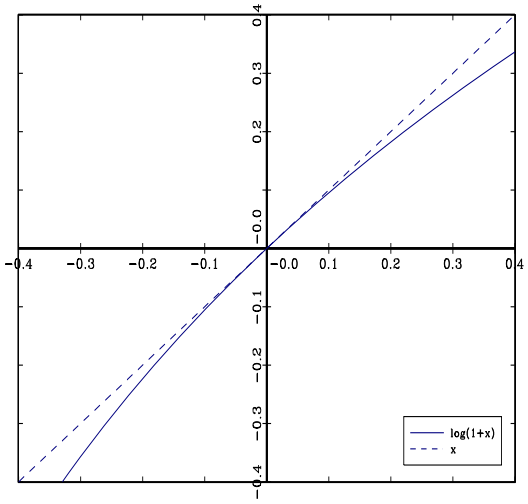


Figure 2.4: $\log(1+x)$

2.5 Conditional Expectation Function

An important determinant of wage levels is education. In many empirical studies economists measure educational attainment by the number of years[8] of schooling, and we will write this variable as *education*.

The conditional mean of log wages given *sex*, *race*, and *education* is a single number for each category. For example

$$\mathbb{E}\left(\log(wage) \mid sex = man,\ race = white,\ education = 12\right) = 2.84.$$

We display in Figure 2.5 the conditional means of $\log(wage)$ for white men and white women as a function of *education*. The plot is quite revealing. We see that the conditional mean is increasing in years of education, but at a different rate for schooling levels above and below nine years. Another striking feature of Figure 2.5 is that the gap between men and women is roughly constant for all education levels. As the variables are measured in logs this implies a constant average percentage gap between men and women regardless of educational attainment.

In many cases it is convenient to simplify the notation by writing variables using single characters, typically y, x and/or z. It is conventional in econometrics to denote the dependent variable (e.g. $\log(wage)$) by the letter y, a conditioning variable (such as *sex*) by the letter x, and multiple conditioning variables (such as *race*, *education* and *sex*) by the subscripted letters $x_1, x_2, ..., x_k$.

Conditional expectations can be written with the generic notation

$$\mathbb{E}\left(y \mid x_1, x_2, ..., x_k\right) = m(x_1, x_2, ..., x_k).$$

We call this the **conditional expectation function** (CEF). The CEF is a function of $(x_1, x_2, ..., x_k)$ as it varies with the variables. For example, the conditional expectation of $y = \log(wage)$ given $(x_1, x_2) = (sex,\ race)$ is given by the six entries of Table 2.1. The CEF is a function of $(sex,\ race)$ as it varies across the entries.

For greater compactness, we will typically write the conditioning variables as a vector in $\mathbb{R}^k$:

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{pmatrix}. \tag{2.5}$$

[8] Here, *education* is defined as years of schooling beyond kindergarten. A high school graduate has *education*=12, a college graduate has *education*=16, a Master's degree has *education*=18, and a professional degree (medical, law or PhD) has *education*=20.

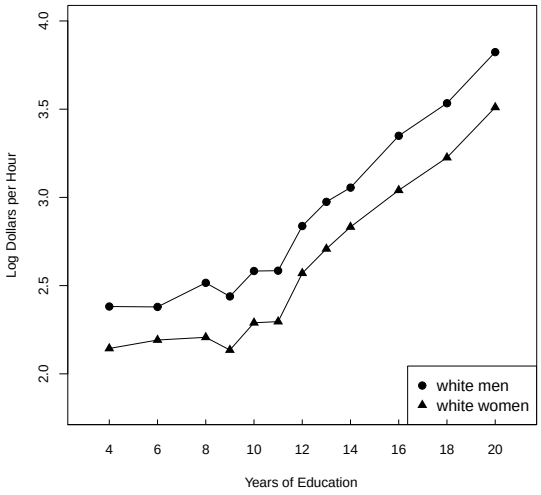


Figure 2.5: Mean Log Wage as a Function of Years of Education

Here we follow the convention of using lower case bold italics $\boldsymbol{x}$ to denote a vector. Given this notation, the CEF can be compactly written as

$$\mathbb{E}(y \mid \boldsymbol{x}) = m(\boldsymbol{x}).$$

The CEF $\mathbb{E}(y \mid \boldsymbol{x})$ is a random variable as it is a function of the random variable $\boldsymbol{x}$. It is also sometimes useful to view the CEF as a function of $\boldsymbol{x}$. In this case we can write $m(\boldsymbol{u}) = \mathbb{E}(y \mid \boldsymbol{x} = \boldsymbol{u})$, which is a function of the argument $\boldsymbol{u}$. The expression $\mathbb{E}(y \mid \boldsymbol{x} = \boldsymbol{u})$ is the conditional expectation of y, given that we know that the random variable $\boldsymbol{x}$ equals the specific value $\boldsymbol{u}$. However, sometimes in econometrics we take a notational shortcut and use $\mathbb{E}(y \mid \boldsymbol{x})$ to refer to this function. Hopefully, the use of $\mathbb{E}(y \mid \boldsymbol{x})$ should be apparent from the context.

2.6 Continuous Variables

In the previous sections, we implicitly assumed that the conditioning variables are discrete. However, many conditioning variables are continuous. In this section, we take up this case and assume that the variables $(y, \boldsymbol{x})$ are continuously distributed with a joint density function $f(y, \boldsymbol{x})$.

As an example, take $y = \log(wage)$ and $x = experience$, the number of years of potential labor market experience[9]. The contours of their joint density are plotted on the left side of Figure 2.6 for the population of white men with 12 years of education.

Given the joint density $f(y, \boldsymbol{x})$ the variable $\boldsymbol{x}$ has the marginal density

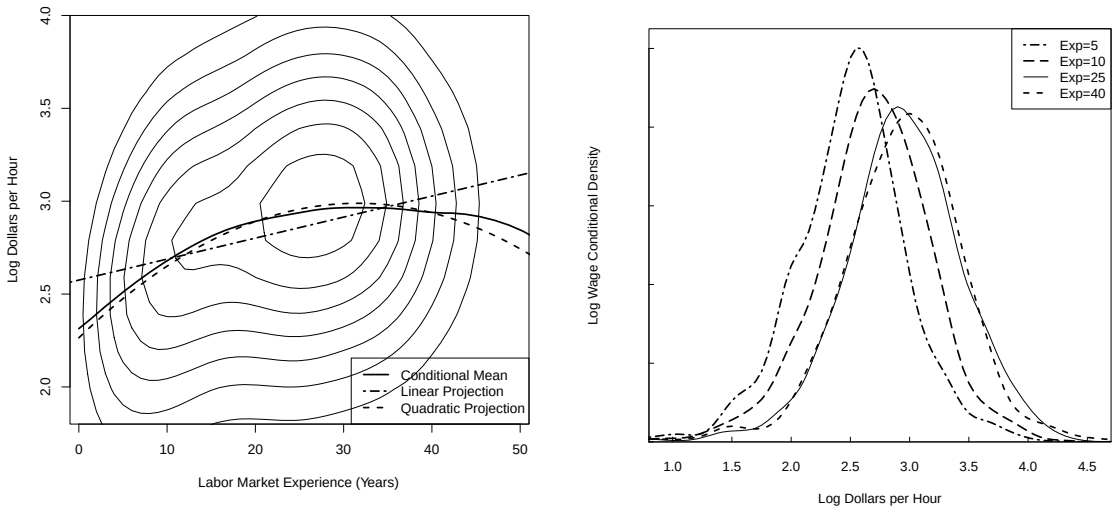
$$f_{\boldsymbol{x}}(\boldsymbol{x}) = \int_{-\infty}^{\infty} f(y, \boldsymbol{x}) dy.$$

For any $\boldsymbol{x}$ such that $f_{\boldsymbol{x}}(\boldsymbol{x}) > 0$ the conditional density of y given $\boldsymbol{x}$ is defined as

$$f_{y|\boldsymbol{x}}(y \mid \boldsymbol{x}) = \frac{f(y, \boldsymbol{x})}{f_{\boldsymbol{x}}(\boldsymbol{x})}. \tag{2.6}$$

The conditional density is a (renormalized) slice of the joint density $f(y, \boldsymbol{x})$ holding $\boldsymbol{x}$ fixed. The slice is renormalized (divided by $f_{\boldsymbol{x}}(\boldsymbol{x})$ so that it integrates to one and is thus a density.) We can

[9] Here, *experience* is defined as potential labor market experience, equal to $age - education - 6$



(a) Joint density of log(wage) and experience and conditional mean

(b) Conditional density

Figure 2.6: White men with *education*=12

visualize this by slicing the joint density function at a specific value of $\boldsymbol{x}$ parallel with the y-axis. For example, take the density contours on the left side of Figure 2.6 and slice through the contour plot at a specific value of *experience*, and then renormalize the slice so that it is a proper density. This gives us the conditional density of log(*wage*) for white men with 12 years of education and this level of experience. We do this for four levels of *experience* (5, 10, 25, and 40 years), and plot these densities on the right side of Figure 2.6. We can see that the distribution of wages shifts to the right and becomes more diffuse as experience increases from 5 to 10 years, and from 10 to 25 years, but there is little change from 25 to 40 years experience.

The CEF of y given $\boldsymbol{x}$ is the mean of the conditional density (2.6)

$$m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x}) = \int_{-\infty}^{\infty} y f_{y|\boldsymbol{x}}(y \mid \boldsymbol{x})\, dy. \tag{2.7}$$

Intuitively, $m(\boldsymbol{x})$ is the mean of y for the idealized subpopulation where the conditioning variables are fixed at $\boldsymbol{x}$. This is idealized since $\boldsymbol{x}$ is continuously distributed so this subpopulation is infinitely small.

This definition (2.7) is appropriate when the conditional density (2.6) is well defined. However, the conditional mean $m(x)$ exists quite generally. In Theorem 2.32.1 in Section 2.32 we show that $m(x)$ exists so long as $\mathbb{E}|y| < \infty$.

In Figure 2.6 the CEF of log(*wage*) given *experience* is plotted as the solid line. We can see that the CEF is a smooth but nonlinear function. The CEF is initially increasing in *experience*, flattens out around *experience* = 30, and then decreases for high levels of experience.

2.7 Law of Iterated Expectations

An extremely useful tool from probability theory is the **law of iterated expectations**. An important special case is the known as the Simple Law.

Theorem 2.7.1 ***Simple Law of Iterated Expectations***
If $\mathbb{E}|y| < \infty$ *then for any random vector* $\boldsymbol{x}$*,*

$$\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x})) = \mathbb{E}(y)$$

The simple law states that the expectation of the conditional expectation is the unconditional expectation. In other words, the average of the conditional averages is the unconditional average. When $\boldsymbol{x}$ is discrete

$$\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x})) = \sum_{j=1}^{\infty} \mathbb{E}(y \mid \boldsymbol{x}_j) \Pr(\boldsymbol{x} = \boldsymbol{x}_j)$$

and when $\boldsymbol{x}$ is continuous

$$\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x})) = \int_{\mathbb{R}^k} \mathbb{E}(y \mid \boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}.$$

Going back to our investigation of average log wages for men and women, the simple law states that

$$\begin{aligned}
&\mathbb{E}(\log(wage) \mid sex = man) \Pr(sex = man) \\
&+ \mathbb{E}(\log(wage) \mid sex = woman) \Pr(sex = woman) \\
&= \mathbb{E}(\log(wage)).
\end{aligned}$$

Or numerically,

$$3.05 \times 0.57 + 2.79 \times 0.43 = 2.92.$$

The general law of iterated expectations allows two sets of conditioning variables.

Theorem 2.7.2 ***Law of Iterated Expectations***
If $\mathbb{E}|y| < \infty$ *then for any random vectors* $\boldsymbol{x}_1$ *and* $\boldsymbol{x}_2$*,*

$$\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) \mid \boldsymbol{x}_1) = \mathbb{E}(y \mid \boldsymbol{x}_1)$$

Notice the way the law is applied. The inner expectation conditions on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, while the outer expectation conditions only on $\boldsymbol{x}_1$. The iterated expectation yields the simple answer $\mathbb{E}(y \mid \boldsymbol{x}_1)$, the expectation conditional on $\boldsymbol{x}_1$ alone. Sometimes we phrase this as: "The smaller information set wins."

As an example

$$\begin{aligned}
&\mathbb{E}(\log(wage) \mid sex = man,\ race = white) \Pr(race = white | sex = man) \\
&+ \mathbb{E}(\log(wage) \mid sex = man,\ race = black) \Pr(race = black | sex = man) \\
&+ \mathbb{E}(\log(wage) \mid sex = man,\ race = other) \Pr(race = other | sex = man) \\
&= \mathbb{E}(\log(wage) \mid sex = man)
\end{aligned}$$

or numerically

$$3.07 \times 0.84 + 2.86 \times 0.08 + 3.03 \times 0.08 = 3.05.$$

A property of conditional expectations is that when you condition on a random vector $\boldsymbol{x}$ you can effectively treat it as if it is constant. For example, $\mathbb{E}(\boldsymbol{x} \mid \boldsymbol{x}) = \boldsymbol{x}$ and $\mathbb{E}(g(\boldsymbol{x}) \mid \boldsymbol{x}) = g(\boldsymbol{x})$ for any function $g(\cdot)$. The general property is known as the Conditioning Theorem.

Theorem 2.7.3 ***Conditioning Theorem***
If $\mathbb{E}|y| < \infty$ *then*

$$\mathbb{E}(g(\boldsymbol{x})\, y \mid \boldsymbol{x}) = g(\boldsymbol{x})\, \mathbb{E}(y \mid \boldsymbol{x}). \tag{2.8}$$

In in addition

$$\mathbb{E}|g(\boldsymbol{x})\, y| < \infty \tag{2.9}$$

then

$$\mathbb{E}(g(\boldsymbol{x})\, y) = \mathbb{E}(g(\boldsymbol{x})\, \mathbb{E}(y \mid \boldsymbol{x})). \tag{2.10}$$

The proofs of Theorems 2.7.1, 2.7.2 and 2.7.3 are given in Section 2.34.

2.8 CEF Error

The CEF error e is defined as the difference between y and the CEF evaluated at the random vector $\boldsymbol{x}$:

$$e = y - m(\boldsymbol{x}).$$

By construction, this yields the formula

$$y = m(\boldsymbol{x}) + e. \tag{2.11}$$

In (2.11) it is useful to understand that the error e is derived from the joint distribution of $(y, \boldsymbol{x})$, and so its properties are derived from this construction.

A key property of the CEF error is that it has a conditional mean of zero. To see this, by the linearity of expectations, the definition $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ and the Conditioning Theorem

$$\begin{aligned}
\mathbb{E}(e \mid \boldsymbol{x}) &= \mathbb{E}((y - m(\boldsymbol{x})) \mid \boldsymbol{x}) \\
&= \mathbb{E}(y \mid \boldsymbol{x}) - \mathbb{E}(m(\boldsymbol{x}) \mid \boldsymbol{x}) \\
&= m(\boldsymbol{x}) - m(\boldsymbol{x}) \\
&= 0.
\end{aligned}$$

This fact can be combined with the law of iterated expectations to show that the unconditional mean is also zero.

$$\mathbb{E}(e) = \mathbb{E}(\mathbb{E}(e \mid \boldsymbol{x})) = \mathbb{E}(0) = 0.$$

We state this and some other results formally.

Theorem 2.8.1 ***Properties of the CEF error***
If $\mathbb{E}|y| < \infty$ *then*

1. $\mathbb{E}(e \mid \boldsymbol{x}) = 0$.
2. $\mathbb{E}(e) = 0$.
3. *If* $\mathbb{E}|y|^r < \infty$ *for* $r \geq 1$ *then* $\mathbb{E}|e|^r < \infty$.
4. *For any function* $h(\boldsymbol{x})$ *such that* $\mathbb{E}|h(\boldsymbol{x})\, e| < \infty$ *then* $\mathbb{E}(h(\boldsymbol{x})\, e) = 0$.

The proof of the third result is deferred to Section 2.34.

The fourth result, whose proof is left to Exercise 2.3, implies that e is uncorrelated with any function of the regressors.

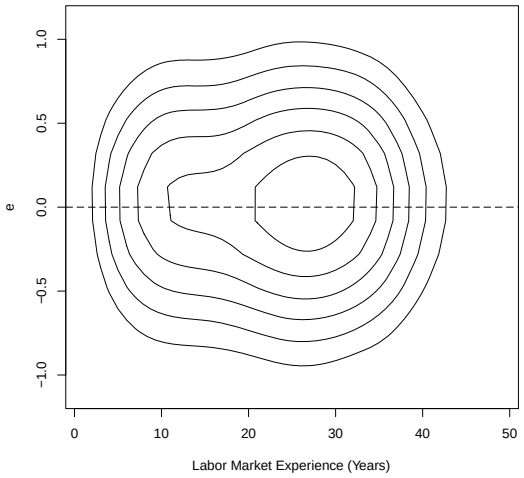


Figure 2.7: Joint density of CEF error e and *experience* for white men with *education*=12.

The equations

$$\begin{aligned} y &= m(\boldsymbol{x}) + e \\ \mathbb{E}\left(e \mid \boldsymbol{x}\right) &= 0 \end{aligned}$$

together imply that $m(\boldsymbol{x})$ is the CEF of y given $\boldsymbol{x}$. It is important to understand that this is not a restriction. These equations hold true by definition.

The condition $\mathbb{E}\left(e \mid \boldsymbol{x}\right) = 0$ is implied by the definition of e as the difference between y and the CEF $m\left(\boldsymbol{x}\right)$. The equation $\mathbb{E}\left(e \mid \boldsymbol{x}\right) = 0$ is sometimes called a conditional mean restriction, since the conditional mean of the error e is restricted to equal zero. The property is also sometimes called **mean independence**, for the conditional mean of e is 0 and thus independent of $\boldsymbol{x}$. However, it does not imply that the distribution of e is independent of $\boldsymbol{x}$. Sometimes the assumption "e is independent of $\boldsymbol{x}$" is added as a convenient simplification, but it is not generic feature of the conditional mean. Typically and generally, e and $\boldsymbol{x}$ are jointly dependent, even though the conditional mean of e is zero.

As an example, the contours of the joint density of e and *experience* are plotted in Figure 2.7 for the same population as Figure 2.6. The error e has a conditional mean of zero for all values of *experience*, but the shape of the conditional distribution varies with the level of *experience*.

As a simple example of a case where x and e are mean independent yet dependent, let $e = x\varepsilon$ where x and ε are independent $\mathrm{N}(0, 1)$. Then conditional on x, the error e has the distribution $\mathrm{N}(0, x^2)$. Thus $\mathbb{E}\left(e \mid x\right) = 0$ and e is mean independent of x, yet e is not fully independent of x. Mean independence does not imply full independence.

2.9 Intercept-Only Model

A special case of the regression model is when there are no regressors $\boldsymbol{x}$. In this case $m(\boldsymbol{x}) = \mathbb{E}\left(y\right) = \mu$, the unconditional mean of y. We can still write an equation for y in the regression format:

$$\begin{aligned} y &= \mu + e \\ \mathbb{E}\left(e\right) &= 0. \end{aligned}$$

This is useful for it unifies the notation.

2.10 Regression Variance

An important measure of the dispersion about the CEF function is the unconditional variance of the CEF error e. We write this as

$$\sigma^2 = \operatorname{var}(e) = \mathbb{E}\left((e - \mathbb{E}e)^2\right) = \mathbb{E}\left(e^2\right).$$

Theorem 2.8.1.3 implies the following simple but useful result.

Theorem 2.10.1 *If* $\mathbb{E}\left(y^2\right) < \infty$ *then* $\sigma^2 < \infty$.

We can call σ^2 the regression variance or the variance of the regression error. The magnitude of σ^2 measures the amount of variation in y which is not "explained" or accounted for in the conditional mean $\mathbb{E}(y \mid \boldsymbol{x})$.

The regression variance depends on the regressors $\boldsymbol{x}$. Consider two regressions

$$y = \mathbb{E}(y \mid \boldsymbol{x}_1) + e_1$$
$$y = \mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) + e_2.$$

We write the two errors distinctly as e_1 and e_2 as they are different – changing the conditioning information changes the conditional mean and therefore the regression error as well.

In our discussion of iterated expectations, we have seen that by increasing the conditioning set, the conditional expectation reveals greater detail about the distribution of y. What is the implication for the regression error?

It turns out that there is a simple relationship. We can think of the conditional mean $\mathbb{E}(y \mid \boldsymbol{x})$ as the "explained portion" of y. The remainder $e = y - \mathbb{E}(y \mid \boldsymbol{x})$ is the "unexplained portion". The simple relationship we now derive shows that the variance of this unexplained portion decreases when we condition on more variables. This relationship is monotonic in the sense that increasing the amont of information always decreases the variance of the unexplained portion.

Theorem 2.10.2 *If* $\mathbb{E}\left(y^2\right) < \infty$ *then*

$$\operatorname{var}(y) \geq \operatorname{var}(y - \mathbb{E}(y \mid \boldsymbol{x}_1)) \geq \operatorname{var}(y - \mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)).$$

Theorem 2.10.2 says that the variance of the difference between y and its conditional mean (weakly) decreases whenever an additional variable is added to the conditioning information.

The proof of Theorem 2.10.2 is given in Section 2.34.

2.11 Best Predictor

Suppose that given a realized value of $\boldsymbol{x}$, we want to create a prediction or forecast of y. We can write any predictor as a function $g(\boldsymbol{x})$ of $\boldsymbol{x}$. The prediction error is the realized difference $y - g(\boldsymbol{x})$. A non-stochastic measure of the magnitude of the prediction error is the expectation of its square

$$\mathbb{E}\left((y - g(\boldsymbol{x}))^2\right). \tag{2.12}$$

We can define the best predictor as the function $g(\boldsymbol{x})$ which minimizes (2.12). What function is the best predictor? It turns out that the answer is the CEF $m(\boldsymbol{x})$. This holds regardless of the joint distribution of $(y, \boldsymbol{x})$.

To see this, note that the mean squared error of a predictor $g(\boldsymbol{x})$ is

$$\begin{aligned}\mathbb{E}\left((y-g(\boldsymbol{x}))^2\right) &= \mathbb{E}\left((e+m(\boldsymbol{x})-g(\boldsymbol{x}))^2\right) \\ &= \mathbb{E}\left(e^2\right)+2\mathbb{E}\left(e\left(m(\boldsymbol{x})-g(\boldsymbol{x})\right)\right)+\mathbb{E}\left(\left(m(\boldsymbol{x})-g(\boldsymbol{x})\right)^2\right) \\ &= \mathbb{E}\left(e^2\right)+\mathbb{E}\left(\left(m(\boldsymbol{x})-g(\boldsymbol{x})\right)^2\right) \\ &\geq \mathbb{E}\left(e^2\right) \\ &= \mathbb{E}\left((y-m(\boldsymbol{x}))^2\right)\end{aligned}$$

where the first equality makes the substitution $y = m(\boldsymbol{x}) + e$ and the third equality uses Theorem 2.8.1.4. The right-hand-side after the third equality is minimized by setting $g(\boldsymbol{x}) = m(\boldsymbol{x})$, yielding the inequality in the fourth line. The minimum is finite under the assumption $\mathbb{E}\left(y^2\right) < \infty$ as shown by Theorem 2.10.1.

We state this formally in the following result.

Theorem 2.11.1 ***Conditional Mean as Best Predictor***
If $\mathbb{E}\left(y^2\right) < \infty$, *then for any predictor* $g(\boldsymbol{x})$,

$$\mathbb{E}\left((y-g(\boldsymbol{x}))^2\right) \geq \mathbb{E}\left((y-m(\boldsymbol{x}))^2\right)$$

where $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$.

It may be helpful to consider this result in the context of the intercept-only model

$$\begin{aligned}y &= \mu + e \\ \mathbb{E}(e) &= 0.\end{aligned}$$

Theorem 2.11.1 shows that the best predictor for y (in the class of constants) is the unconditional mean $\mu = \mathbb{E}(y)$, in the sense that the mean minimizes the mean squared prediction error.

2.12 Conditional Variance

While the conditional mean is a good measure of the location of a conditional distribution, it does not provide information about the spread of the distribution. A common measure of the dispersion is the **conditional variance**. We first give the general definition of the conditional variance of a random variable w.

Definition 2.12.1 *If* $\mathbb{E}\left(w^2\right) < \infty$, *the* ***conditional variance*** *of* w *given* $\boldsymbol{x}$ *is*

$$\operatorname{var}(w \mid \boldsymbol{x}) = \mathbb{E}\left((w-\mathbb{E}(w \mid \boldsymbol{x}))^2 \mid \boldsymbol{x}\right)$$

Notice that the conditional variance is the conditional second moment, centered around the conditional first moment. Given this definition, we define the conditional variance of the regression error.

Definition 2.12.2 *If* $\mathbb{E}\left(e^2\right) < \infty$, *the* ***conditional variance*** *of the regression error* e *is*

$$\sigma^2(\boldsymbol{x}) = \operatorname{var}\left(e \mid \boldsymbol{x}\right) = \mathbb{E}\left(e^2 \mid \boldsymbol{x}\right).$$

Generally, $\sigma^2\left(\boldsymbol{x}\right)$ is a non-trivial function of $\boldsymbol{x}$ and can take any form subject to the restriction that it is non-negative. One way to think about $\sigma^2(\boldsymbol{x})$ is that it is the conditional mean of e^2 given $\boldsymbol{x}$. Notice as well that $\sigma^2(\boldsymbol{x}) = \operatorname{var}\left(y \mid \boldsymbol{x}\right)$ so it is equivalently the conditional variance of the dependent variable.

The variance is in a different unit of measurement than the original variable. To convert the variance back to the same unit of measure we define the **conditional standard deviation** as its square root $\sigma(\boldsymbol{x}) = \sqrt{\sigma^2(\boldsymbol{x})}$.

As an example of how the conditional variance depends on observables, compare the conditional log wage densities for men and women displayed in Figure 2.3. The difference between the densities is not purely a location shift, but is also a difference in spread. Specifically, we can see that the density for men's log wages is somewhat more spread out than that for women, while the density for women's wages is somewhat more peaked. Indeed, the conditional standard deviation for men's wages is 3.05 and that for women is 2.81. So while men have higher average wages, they are also somewhat more dispersed.

The unconditional error variance and the conditional variance are related by the law of iterated expectations

$$\sigma^2 = \mathbb{E}\left(e^2\right) = \mathbb{E}\left(\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right)\right) = \mathbb{E}\left(\sigma^2(\boldsymbol{x})\right).$$

That is, the unconditional error variance is the average conditional variance.

Given the conditional variance, we can define a rescaled error

$$\varepsilon = \frac{e}{\sigma(\boldsymbol{x})}. \tag{2.13}$$

We can calculate that since $\sigma(\boldsymbol{x})$ is a function of $\boldsymbol{x}$

$$\mathbb{E}\left(\varepsilon \mid \boldsymbol{x}\right) = \mathbb{E}\left(\frac{e}{\sigma(\boldsymbol{x})} \mid \boldsymbol{x}\right) = \frac{1}{\sigma(\boldsymbol{x})}\mathbb{E}\left(e \mid \boldsymbol{x}\right) = 0$$

and

$$\operatorname{var}\left(\varepsilon \mid \boldsymbol{x}\right) = \mathbb{E}\left(\varepsilon^2 \mid \boldsymbol{x}\right) = \mathbb{E}\left(\frac{e^2}{\sigma^2(\boldsymbol{x})} \mid \boldsymbol{x}\right) = \frac{1}{\sigma^2(\boldsymbol{x})}\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \frac{\sigma^2(\boldsymbol{x})}{\sigma^2(\boldsymbol{x})} = 1.$$

Thus ε has a conditional mean of zero, and a conditional variance of 1.

Notice that (2.13) can be rewritten as

$$e = \sigma(\boldsymbol{x})\varepsilon.$$

and substituting this for e in the CEF equation (2.11), we find that

$$y = m(\boldsymbol{x}) + \sigma(\boldsymbol{x})\varepsilon. \tag{2.14}$$

This is an alternative (mean-variance) representation of the CEF equation.

Many econometric studies focus on the conditional mean $m(\boldsymbol{x})$ and either ignore the conditional variance $\sigma^2(\boldsymbol{x})$, treat it as a constant $\sigma^2(\boldsymbol{x}) = \sigma^2$, or treat it as a nuisance parameter (a parameter not of primary interest). This is appropriate when the primary variation in the conditional distribution is in the mean, but can be short-sighted in other cases. Dispersion is relevant to many economic topics, including income and wealth distribution, economic inequality, and price dispersion. Conditional dispersion (variance) can be a fruitful subject for investigation.

The perverse consequences of a narrow-minded focus on the mean has been parodied in a classic joke:

> An economist was standing with one foot in a bucket of boiling water and the other foot in a bucket of ice. When asked how he felt, he replied, "On average I feel just fine."

Clearly, the economist in question ignored variance!

2.13 Homoskedasticity and Heteroskedasticity

An important special case obtains when the conditional variance $\sigma^2(\boldsymbol{x})$ is a constant and independent of $\boldsymbol{x}$. This is called **homoskedasticity**.

Definition 2.13.1 *The error is* ***homoskedastic*** *if* $\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \sigma^2$ *does not depend on* $\boldsymbol{x}$.

In the general case where $\sigma^2(\boldsymbol{x})$ depends on $\boldsymbol{x}$ we say that the error e is **heteroskedastic**.

Definition 2.13.2 *The error is* ***heteroskedastic*** *if* $\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \sigma^2(\boldsymbol{x})$ *depends on* $\boldsymbol{x}$.

It is helpful to understand that the concepts homoskedasticity and heteroskedasticity concern the conditional variance, not the unconditional variance. By definition, the unconditional variance σ^2 is a constant and independent of the regressors $\boldsymbol{x}$. So when we talk about the variance as a function of the regressors, we are talking about the conditional variance $\sigma^2(\boldsymbol{x})$.

Some older or introductory textbooks describe heteroskedasticity as the case where "the variance of e varies across observations". This is a poor and confusing definition. It is more constructive to understand that heteroskedasticity means that the conditional variance $\sigma^2(\boldsymbol{x})$ depends on observables.

Older textbooks also tend to describe homoskedasticity as a component of a correct regression specification, and describe heteroskedasticity as an exception or deviance. This description has influenced many generations of economists, but it is unfortunately backwards. The correct view is that heteroskedasticity is generic and "standard", while homoskedasticity is unusual and exceptional. The default in empirical work should be to assume that the errors are heteroskedastic, not the converse.

In apparent contradiction to the above statement, we will still frequently impose the homoskedasticity assumption when making theoretical investigations into the properties of estimation and inference methods. The reason is that in many cases homoskedasticity greatly simplifies the theoretical calculations, and it is therefore quite advantageous for teaching and learning. It should always be remembered, however, that homoskedasticity is never imposed because it is believed to be a correct feature of an empirical model, but rather because of its simplicity.

2.14 Regression Derivative

One way to interpret the CEF $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ is in terms of how marginal changes in the regressors $\boldsymbol{x}$ imply changes in the conditional mean of the response variable y. It is typical to consider marginal changes in a single regressor, say x_1, holding the remainder fixed. When a regressor x_1 is continuously distributed, we define the marginal effect of a change in x_1, holding the variables $x_2, ..., x_k$ fixed, as the partial derivative of the CEF

$$\frac{\partial}{\partial x_1} m(x_1, ..., x_k).$$

When x_1 is discrete we define the marginal effect as a discrete difference. For example, if x_1 is binary, then the marginal effect of x_1 on the CEF is

$$m(1, x_2, ..., x_k) - m(0, x_2, ..., x_k).$$

We can unify the continuous and discrete cases with the notation

$$\nabla_1 m(\boldsymbol{x}) = \begin{cases} \dfrac{\partial}{\partial x_1} m(x_1, ..., x_k), & \text{if } x_1 \text{ is continuous} \\ \\ m(1, x_2, ..., x_k) - m(0, x_2, ..., x_k), & \text{if } x_1 \text{ is binary.} \end{cases}$$

Collecting the k effects into one $k \times 1$ vector, we define the **regression derivative** with respect to $\boldsymbol{x}$:

$$\boldsymbol{\nabla} m(\boldsymbol{x}) = \begin{bmatrix} \nabla_1 m(\boldsymbol{x}) \\ \nabla_2 m(\boldsymbol{x}) \\ \vdots \\ \nabla_k m(\boldsymbol{x}) \end{bmatrix}.$$

When all elements of $\boldsymbol{x}$ are continuous, then we have the simplification $\boldsymbol{\nabla} m(\boldsymbol{x}) = \dfrac{\partial}{\partial \boldsymbol{x}} m(\boldsymbol{x})$, the vector of partial derivatives.

There are two important points to remember concerning our definition of the regression derivative.

First, the effect of each variable is calculated holding the other variables constant. This is the **ceteris paribus** concept commonly used in economics. But in the case of a regression derivative, the conditional mean does not literally hold *all else* constant. It only holds constant the variables included in the conditional mean. This means that the regression derivative depends on which regressors are included. For example, in a regression of wages on education, experience, race and sex, the regression derivative with respect to education shows the marginal effect of education on mean wages, holding constant experience, race and sex. But it does not hold constant an individual's unobservable characteristics (such as ability), nor variables not included in the regression (such as the quality of education).

Second, the regression derivative is the change in the conditional expectation of y, not the change in the actual value of y for an individual. It is tempting to think of the regression derivative as the change in the actual value of y, but this is not a correct interpretation. The regression derivative $\boldsymbol{\nabla} m(\boldsymbol{x})$ is the change in the actual value of y only if the error e is unaffected by the change in the regressor $\boldsymbol{x}$. We return to a discussion of causal effects in Section 2.29.

2.15 Linear CEF

An important special case is when the CEF $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ is linear in $\boldsymbol{x}$. In this case we can write the mean equation as

$$m(\boldsymbol{x}) = x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k + \beta_{k+1}.$$

Notationally it is convenient to write this as a simple function of the vector $\boldsymbol{x}$. An easy way to do so is to augment the regressor vector $\boldsymbol{x}$ by listing the number "1" as an element. We call this the "constant" and the corresponding coefficient is called the "intercept". Equivalently, specify that the final element[10] of the vector $\boldsymbol{x}$ is $x_k = 1$. Thus (2.5) has been redefined as the $k \times 1$ vector

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{k-1} \\ 1 \end{pmatrix}. \tag{2.15}$$

With this redefinition, the CEF is

$$\begin{aligned} m(\boldsymbol{x}) &= x_1\beta_1 + x_2\beta_2 + \cdots + \beta_k \\ &= \boldsymbol{x}'\boldsymbol{\beta} \end{aligned} \tag{2.16}$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{2.17}$$

is a $k \times 1$ coefficient vector. This is the **linear CEF model**. It is also often called the **linear regression model**, or the regression of y on $\boldsymbol{x}$.

In the linear CEF model, the regression derivative is simply the coefficient vector. That is

$$\boldsymbol{\nabla} m(\boldsymbol{x}) = \boldsymbol{\beta}.$$

This is one of the appealing features of the linear CEF model. The coefficients have simple and natural interpretations as the marginal effects of changing one variable, holding the others constant.

Linear CEF Model

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ \mathbb{E}\left(e \mid \boldsymbol{x}\right) &= 0 \end{aligned}$$

If in addition the error is homoskedastic, we call this the homoskedastic linear CEF model.

Homoskedastic Linear CEF Model

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ \mathbb{E}\left(e \mid \boldsymbol{x}\right) &= 0 \\ \mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) &= \sigma^2 \end{aligned}$$

[10] The order doesn't matter. It could be any element.

2.16 Linear CEF with Nonlinear Effects

The linear CEF model of the previous section is less restrictive than it might appear, as we can include as regressors nonlinear transformations of the original variables. In this sense, the linear CEF framework is flexible and can capture many nonlinear effects.

For example, suppose we have two scalar variables x_1 and x_2. The CEF could take the quadratic form

$$m(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6. \tag{2.18}$$

This equation is quadratic in the regressors (x_1, x_2) yet linear in the coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_6)'$. We will descriptively call (2.18) a quadratic CEF, and yet (2.18) is also a linear CEF in the sense of being linear in the coefficients. The key is to understand that (2.18) is quadratic in the variables (x_1, x_2) yet linear in the coefficients $\boldsymbol{\beta}$.

To simplify the expression, we define the transformations $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1x_2$, and $x_6 = 1$, and redefine the regressor vector as $\boldsymbol{x} = (x_1, ..., x_6)'$. With this redefinition,

$$m(x_1, x_2) = \boldsymbol{x}'\boldsymbol{\beta}$$

which is linear in $\boldsymbol{\beta}$. For most econometric purposes (estimation and inference on $\boldsymbol{\beta}$) the linearity in $\boldsymbol{\beta}$ is all that is important.

An exception is in the analysis of regression derivatives. In nonlinear equations such as (2.18), the regression derivative should be defined with respect to the original variables, not with respect to the transformed variables. Thus

$$\frac{\partial}{\partial x_1} m(x_1, x_2) = \beta_1 + 2x_1\beta_3 + x_2\beta_5$$
$$\frac{\partial}{\partial x_2} m(x_1, x_2) = \beta_2 + 2x_2\beta_4 + x_1\beta_5.$$

We see that in the model (2.18), the regression derivatives are not a simple coefficient, but are functions of several coefficients plus the levels of (x_1,x_2). Consequently it is difficult to interpret the coefficients individually. It is more useful to interpret them as a group.

We typically call β_5 the **interaction effect**. Notice that it appears in both regression derivative equations, and has a symmetric interpretation in each. If $\beta_5 > 0$ then the regression derivative with respect to x_1 is increasing in the level of x_2 (and the regression derivative with respect to x_2 is increasing in the level of x_1), while if $\beta_5 < 0$ the reverse is true.

2.17 Linear CEF with Dummy Variables

When all regressors take a finite set of values, it turns out the CEF can be written as a linear function of regressors.

This simplest example is a **binary** variable, which takes only two distinct values. For example, in most data sets the variable *sex* takes only the values *man* and *woman* (or male and female). Binary variables are extremely common in econometric applications, and are alternatively called **dummy variables** or **indicator variables**.

Consider the simple case of a single binary regressor. In this case, the conditional mean can only take two distinct values. For example,

$$\mathbb{E}(y \mid sex) = \begin{cases} \mu_0 & \text{if} \quad sex{=}man \\ \\ \mu_1 & \text{if} \quad sex{=}woman \end{cases}.$$

To facilitate a mathematical treatment, we typically record dummy variables with the values $\{0, 1\}$. For example

$$x_1 = \begin{cases} 0 & \text{if} \quad sex{=}man \\ 1 & \text{if} \quad sex{=}woman \end{cases}. \tag{2.19}$$

Given this notation we can write the conditional mean as a linear function of the dummy variable x_1, that is

$$\mathbb{E}(y \mid x_1) = \beta_1 x_1 + \beta_2$$

where $\beta_1 = \mu_1 - \mu_0$ and $\beta_2 = \mu_0$. In this simple regression equation the intercept β_2 is equal to the conditional mean of y for the $x_1 = 0$ subpopulation (men) and the slope β_1 is equal to the difference in the conditional means between the two subpopulations.

Equivalently, we could have defined x_1 as

$$x_1 = \begin{cases} 1 & \text{if} \quad \textit{sex=man} \\ 0 & \text{if} \quad \textit{sex=woman} \end{cases}. \tag{2.20}$$

In this case, the regression intercept is the mean for women (rather than for men) and the regression slope has switched signs. The two regressions are equivalent but the interpretation of the coefficients has changed. Therefore it is always important to understand the precise definitions of the variables, and illuminating labels are helpful. For example, labelling x_1 as "sex" does not help distinguish between definitions (2.19) and (2.20). Instead, it is better to label x_1 as "women" or "female" if definition (2.19) is used, or as "men" or "male" if (2.20) is used.

Now suppose we have two dummy variables x_1 and x_2. For example, $x_2 = 1$ if the person is married, else $x_2 = 0$. The conditional mean given x_1 and x_2 takes at most four possible values:

$$\mathbb{E}(y \mid x_1, x_2) = \begin{cases} \mu_{00} & \text{if} \quad x_1 = 0 \text{ and } x_2 = 0 & \text{(unmarried men)} \\ \mu_{01} & \text{if} \quad x_1 = 0 \text{ and } x_2 = 1 & \text{(married men)} \\ \mu_{10} & \text{if} \quad x_1 = 1 \text{ and } x_2 = 0 & \text{(unmarried women)} \\ \mu_{11} & \text{if} \quad x_1 = 1 \text{ and } x_2 = 1 & \text{(married women)} \end{cases}.$$

In this case we can write the conditional mean as a linear function of x_1, x_2 and their product $x_1 x_2$:

$$\mathbb{E}(y \mid x_1, x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4$$

where $\beta_1 = \mu_{10} - \mu_{00}$, $\beta_2 = \mu_{01} - \mu_{00}$, $\beta_3 = \mu_{11} - \mu_{10} - \mu_{01} + \mu_{00}$, and $\beta_4 = \mu_{00}$.

We can view the coefficient β_1 as the effect of sex on expected log wages for unmarried wage earners, the coefficient β_2 as the effect of marriage on expected log wages for men wage earners, and the coefficient β_3 as the difference between the effects of marriage on expected log wages among women and among men. Alternatively, it can also be interpreted as the difference between the effects of sex on expected log wages among married and non-married wage earners. Both interpretations are equally valid. We often describe β_3 as measuring the **interaction** between the two dummy variables, or the **interaction effect**, and describe $\beta_3 = 0$ as the case when the interaction effect is zero.

In this setting we can see that the CEF is linear in the three variables $(x_1, x_2, x_1 x_2)$. Thus to put the model in the framework of Section 2.15, we would define the regressor $x_3 = x_1 x_2$ and the regressor vector as

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix}.$$

So even though we started with only 2 dummy variables, the number of regressors (including the intercept) is 4.

If there are 3 dummy variables x_1, x_2, x_3, then $\mathbb{E}(y \mid x_1, x_2, x_3)$ takes at most $2^3 = 8$ distinct values and can be written as the linear function

$$\mathbb{E}(y \mid x_1, x_2, x_3) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3 + \beta_8$$

which has eight regressors including the intercept.

In general, if there are p dummy variables $x_1, ..., x_p$ then the CEF $\mathbb{E}(y \mid x_1, x_2, ..., x_p)$ takes at most 2^p distinct values, and can be written as a linear function of the 2^p regressors including $x_1, x_2, ..., x_p$ and all cross-products. This might be excessive in practice if p is modestly large. In the next section we will discuss projection approximations which yield more parsimonious parameterizations.

We started this section by saying that the conditional mean is linear whenever all regressors take only a finite number of possible values. How can we see this? Take a **categorical** variable, such as *race*. For example, we earlier divided race into three categories. We can record categorical variables using numbers to indicate each category, for example

$$x_3 = \begin{cases} 1 & \text{if} \quad \textit{white} \\ 2 & \text{if} \quad \textit{black} \\ 3 & \text{if} \quad \textit{other} \end{cases}.$$

When doing so, the values of x_3 have no meaning in terms of magnitude, they simply indicate the relevant category.

When the regressor is categorical the conditional mean of y given x_3 takes a distinct value for each possibility:

$$\mathbb{E}(y \mid x_3) = \begin{cases} \mu_1 & \text{if} \quad x_3 = 1 \\ \mu_2 & \text{if} \quad x_3 = 2 \\ \mu_3 & \text{if} \quad x_3 = 3 \end{cases}.$$

This is not a linear function of x_3 itself, but it can be made a linear function by constructing dummy variables for two of the three categories. For example

$$x_4 = \begin{cases} 1 & \text{if} \quad \textit{black} \\ 0 & \text{if} \quad \textit{not black} \end{cases}$$

$$x_5 = \begin{cases} 1 & \text{if} \quad \textit{other} \\ 0 & \text{if} \quad \textit{not other} \end{cases}.$$

In this case, the categorical variable x_3 is equivalent to the pair of dummy variables (x_4, x_5). The explicit relationship is

$$x_3 = \begin{cases} 1 & \text{if} \quad x_4 = 0 \text{ and } x_5 = 0 \\ 2 & \text{if} \quad x_4 = 1 \text{ and } x_5 = 0 \\ 3 & \text{if} \quad x_4 = 0 \text{ and } x_5 = 1 \end{cases}.$$

Given these transformations, we can write the conditional mean of y as a linear function of x_4 and x_5

$$\mathbb{E}(y \mid x_3) = \mathbb{E}(y \mid x_4, x_5) = \beta_1 x_4 + \beta_2 x_5 + \beta_3.$$

We can write the CEF as either $\mathbb{E}(y \mid x_3)$ or $\mathbb{E}(y \mid x_4, x_5)$ (they are equivalent), but it is only linear as a function of x_4 and x_5.

This setting is similar to the case of two dummy variables, with the difference that we have not included the interaction term $x_4 x_5$. This is because the event $\{x_4 = 1 \text{ and } x_5 = 1\}$ is empty by construction, so $x_4 x_5 = 0$ by definition.

2.18 Best Linear Predictor

While the conditional mean $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ is the best predictor of y among all functions of $\boldsymbol{x}$, its functional form is typically unknown. In particular, the linear CEF model is empirically unlikely to be accurate unless $\boldsymbol{x}$ is discrete and low-dimensional so all interactions are included. Consequently in most cases it is more realistic to view the linear specification (2.16) as an approximation. In this section we derive a specific approximation with a simple interpretation.

Theorem 2.11.1 showed that the conditional mean $m(\boldsymbol{x})$ is the best predictor in the sense that it has the lowest mean squared error among all predictors. By extension, we can define an approximation to the CEF by the linear function with the lowest mean squared error among all linear predictors.

For this derivation we require the following regularity condition.

Assumption 2.18.1

1. $\mathbb{E}\left(y^2\right) < \infty$.
2. $\mathbb{E}\|\boldsymbol{x}\|^2 < \infty$.
3. $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ *is positive definite.*

In Assumption 2.18.1.2 we use the notation $\|\boldsymbol{x}\| = \left(\boldsymbol{x}'\boldsymbol{x}\right)^{1/2}$ to denote the Euclidean length of the vector $\boldsymbol{x}$.

The first two parts of Assumption 2.18.1 imply that the variables y and $\boldsymbol{x}$ have finite means, variances, and covariances. The third part of the assumption is more technical, and its role will become apparent shortly. It is equivalent to imposing that the columns of the matrix $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ are linearly independent, or that the matrix is invertible.

A linear predictor for y is a function of the form $\boldsymbol{x}'\boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^k$. The mean squared prediction error is

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2\right).$$

The **best linear predictor** of y given $\boldsymbol{x}$, written $\mathcal{P}(y \mid \boldsymbol{x})$, is found by selecting the vector $\boldsymbol{\beta}$ to minimize $S(\boldsymbol{\beta})$.

Definition 2.18.1 *The* ***Best Linear Predictor*** *of y given $\boldsymbol{x}$ is*

$$\mathcal{P}(y \mid \boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$$

where $\boldsymbol{\beta}$ minimizes the mean squared prediction error

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2\right).$$

The minimizer

$$\boldsymbol{\beta} = \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^k} S(\boldsymbol{b}) \tag{2.21}$$

is called the ***Linear Projection Coefficient****.*

We now calculate an explicit expression for its value. The mean squared prediction error can be written out as a quadratic function of $\boldsymbol{\beta}$:

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(y^2\right) - 2\boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}y\right) + \boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta}.$$

The quadratic structure of $S(\boldsymbol{\beta})$ means that we can solve explicitly for the minimizer. The first-order condition for minimization (from Appendix A.15) is

$$\mathbf{0} = \frac{\partial}{\partial \boldsymbol{\beta}} S(\boldsymbol{\beta}) = -2\mathbb{E}\left(\boldsymbol{x}y\right) + 2\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta}. \tag{2.22}$$

Rewriting (2.22) as

$$2\mathbb{E}(\boldsymbol{x}y) = 2\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta}$$

and dividing by 2, this equation takes the form

$$\boldsymbol{Q}_{\boldsymbol{x}y} = \boldsymbol{Q}_{\boldsymbol{x}\boldsymbol{x}}\boldsymbol{\beta} \tag{2.23}$$

where $\boldsymbol{Q}_{\boldsymbol{x}y} = \mathbb{E}(\boldsymbol{x}y)$ is $k \times 1$ and $\boldsymbol{Q}_{\boldsymbol{x}\boldsymbol{x}} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ is $k \times k$. The solution is found by inverting the matrix $\boldsymbol{Q}_{\boldsymbol{x}\boldsymbol{x}}$, and is written

$$\boldsymbol{\beta} = \boldsymbol{Q}_{\boldsymbol{x}\boldsymbol{x}}^{-1}\boldsymbol{Q}_{\boldsymbol{x}y}$$

or

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y). \tag{2.24}$$

It is worth taking the time to understand the notation involved in the expression (2.24). $\boldsymbol{Q}_{\boldsymbol{x}\boldsymbol{x}}$ is a $k \times k$ matrix and $\boldsymbol{Q}_{\boldsymbol{x}y}$ is a $k \times 1$ column vector. Therefore, alternative expressions such as $\frac{\mathbb{E}(\boldsymbol{x}y)}{\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')}$ or $\mathbb{E}(\boldsymbol{x}y)\left(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')\right)^{-1}$ are incoherent and incorrect. We also can now see the role of Assumption 2.18.1.3. It is equivalent to assuming that $\boldsymbol{Q}_{\boldsymbol{x}\boldsymbol{x}}$ has an inverse $\boldsymbol{Q}_{\boldsymbol{x}\boldsymbol{x}}^{-1}$ which is necessary for the normal equations (2.23) to have a solution or equivalently for (2.24) to be uniquely defined. In the absence of Assumption 2.18.1.3 there could be multiple solutions to the equation (2.23).

We now have an explicit expression for the best linear predictor:

$$\mathcal{P}(y \mid \boldsymbol{x}) = \boldsymbol{x}'\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y).$$

This expression is also referred to as the **linear projection** of y on $\boldsymbol{x}$.

The **projection error** is

$$e = y - \boldsymbol{x}'\boldsymbol{\beta}. \tag{2.25}$$

This equals the error (2.11) from the regression equation when (and only when) the conditional mean is linear in $\boldsymbol{x}$, otherwise they are distinct.

Rewriting, we obtain a decomposition of y into linear predictor and error

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e. \tag{2.26}$$

In general we call equation (2.26) or $\boldsymbol{x}'\boldsymbol{\beta}$ the best linear predictor of y given $\boldsymbol{x}$**,** or the linear projection of y on $\boldsymbol{x}$. Equation (2.26) is also often called the **regression** of y on $\boldsymbol{x}$ but this can sometimes be confusing as economists use the term *regression* in many contexts. (Recall that we said in Section 2.15 that the linear CEF model is also called the linear regression model.)

An important property of the projection error e is

$$\mathbb{E}(\boldsymbol{x}e) = \boldsymbol{0}. \tag{2.27}$$

To see this, using the definitions (2.25) and (2.24) and the matrix properties $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$ and $\boldsymbol{I}\,\boldsymbol{a} = \boldsymbol{a}$,

$$\begin{aligned}\mathbb{E}(\boldsymbol{x}e) &= \mathbb{E}\left(\boldsymbol{x}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\right) \\ &= \mathbb{E}(\boldsymbol{x}y) - \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y) \\ &= \boldsymbol{0}\end{aligned} \tag{2.28}$$

as claimed.

Equation (2.27) is a set of k equations, one for each regressor. In other words, (2.27) is equivalent to

$$\mathbb{E}(x_j e) = 0 \tag{2.29}$$

for $j = 1, ..., k$. As in (2.15), the regressor vector $\boldsymbol{x}$ typically contains a constant, e.g. $x_k = 1$. In this case (2.29) for $j = k$ is the same as

$$\mathbb{E}(e) = 0. \tag{2.30}$$

Thus the projection error has a mean of zero when the regressor vector contains a constant. (When $\boldsymbol{x}$ does not have a constant, (2.30) is not guaranteed. As it is desirable for e to have a zero mean, this is a good reason to always include a constant in any regression model.)

It is also useful to observe that since $\operatorname{cov}(x_j, e) = \mathbb{E}(x_j e) - \mathbb{E}(x_j)\mathbb{E}(e)$, then (2.29)-(2.30) together imply that the variables x_j and e are uncorrelated.

This completes the derivation of the model. We summarize some of the most important properties.

Theorem 2.18.1 ***Properties of Linear Projection Model***
Under Assumption 2.18.1,

1. *The moments $\mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ and $\mathbb{E}(\boldsymbol{x}y)$ exist with finite elements.*
2. *The Linear Projection Coefficient (2.21) exists, is unique, and equals*
$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y).$$
3. *The best linear predictor of y given $\boldsymbol{x}$ is*
$$\mathcal{P}(y \mid \boldsymbol{x}) = \boldsymbol{x}'\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y).$$
4. *The projection error $e = y - \boldsymbol{x}'\boldsymbol{\beta}$ exists and satisfies*
$$\mathbb{E}\left(e^2\right) < \infty$$
and
$$\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}.$$
5. *If $\boldsymbol{x}$ contains an constant, then*
$$\mathbb{E}(e) = 0.$$
6. *If $\mathbb{E}|y|^r < \infty$ and $\mathbb{E}\|\boldsymbol{x}\|^r < \infty$ for $r \geq 2$ then $\mathbb{E}|e|^r < \infty$.*

A complete proof of Theorem 2.18.1 is given in Section 2.34.

It is useful to reflect on the generality of Theorem 2.18.1. The only restriction is Assumption 2.18.1. Thus for any random variables $(y, \boldsymbol{x})$ with finite variances we can define a linear equation (2.26) with the properties listed in Theorem 2.18.1. Stronger assumptions (such as the linear CEF model) are not necessary. In this sense the linear model (2.26) exists quite generally. However, it is important not to misinterpret the generality of this statement. The linear equation (2.26) is defined as the best linear predictor. It is not necessarily a conditional mean, nor a parameter of a structural or causal economic model.

Linear Projection Model

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e.$$
$$\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}$$
$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y)$$

We illustrate projection using three log wage equations introduced in earlier sections.

For our first example, we consider a model with the two dummy variables for sex and race similar to Table 2.1. As we learned in Section 2.17, the entries in this table can be equivalently expressed by a linear CEF. For simplicity, let's consider the CEF of $\log(wage)$ as a function of *Black* and *Female*.

$$\mathbb{E}(\log(wage) \mid Black, Female) = -0.20Black - 0.24Female + 0.10Black \times Female + 3.06. \quad (2.31)$$

This is a CEF as the variables are binary and all interactions are included.

Now consider a simpler model omitting the interaction effect. This is the linear projection on the variables *Black* and *Female*

$$\mathcal{P}(\log(wage) \mid Black, Female) = -0.15Black - 0.23Female + 3.06. \quad (2.32)$$

What is the difference? The full CEF (2.31) shows that the race gap is differentiated by sex: it is 20% for black men (relative to non-black men) and 10% for black women (relative to non-black women). The projection model (2.32) simplifies this analysis, calculating an average 15% wage gap for blacks, ignoring the role of sex. Notice that this is despite the fact that the sex variable is included in (2.32).

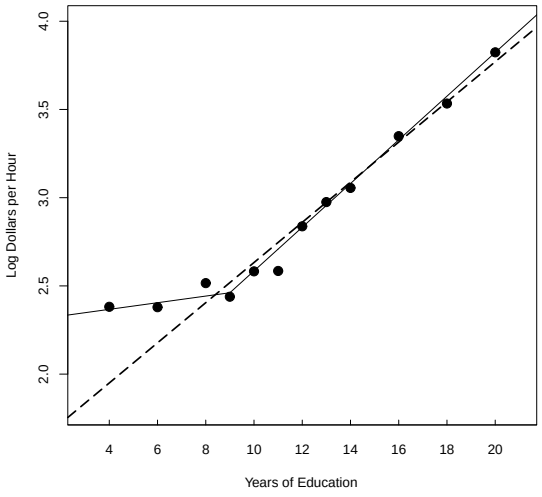


Figure 2.8: Projections of $\log(wage)$ onto Education

For our second example we consider the CEF of log wages as a function of years of education for white men which was illustrated in Figure 2.5 and is repeated in Figure 2.8. Superimposed on the figure are two projections. The first (given by the dashed line) is the linear projection of log wages on years of education

$$\mathcal{P}(\log(wage) \mid Education) = 0.11Education + 1.5.$$

This simple equation indicates an average 11% increase in wages for every year of education. An inspection of the Figure shows that this approximation works well for *education*$\geq$ 9, but under-predicts for individuals with lower levels of education. To correct this imbalance we use a linear spline equation which allows different rates of return above and below 9 years of education:

$$\begin{aligned}&\mathcal{P}\left(\log(wage) \mid Education, (Education - 9) \times 1\left(Education > 9\right)\right)\\ &= 0.02Education + 0.10 \times (Education - 9) \times 1\left(Education > 9\right) + 2.3.\end{aligned}$$

This equation is displayed in Figure 2.8 using the solid line, and appears to fit much better. It indicates a 2% increase in mean wages for every year of education below 9, and a 12% increase in mean wages for every year of education above 9. It is still an approximation to the conditional mean but it appears to be fairly reasonable.

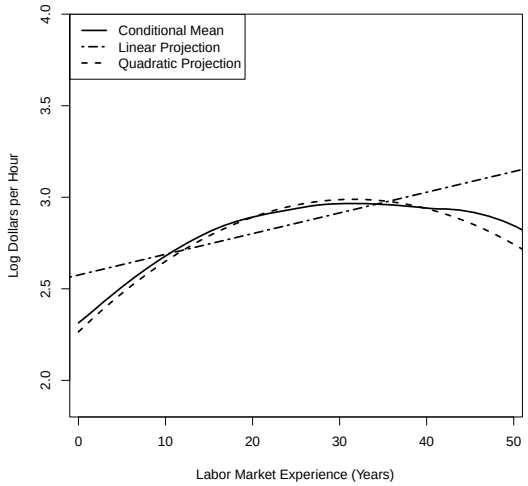


Figure 2.9: Linear and Quadratic Projections of log(*wage*) onto Experience

For our third example we take the CEF of log wages as a function of years of experience for white men with 12 years of education, which was illustrated in Figure 2.6 and is repeated as the solid line in Figure 2.9. Superimposed on the figure are two projections. The first (given by the dot-dashed line) is the linear projection on experience

$$\mathcal{P}(\log(wage) \mid Experience) = 0.011Experience + 2.5$$

and the second (given by the dashed line) is the linear projection on experience and its square

$$\mathcal{P}(\log(wage) \mid Experience) = 0.046Experience - 0.0007Experience^2 + 2.3.$$

It is fairly clear from an examination of Figure 2.9 that the first linear projection is a poor approximation. It over-predicts wages for young and old workers, and under-predicts for the rest. Most importantly, it misses the strong downturn in expected wages for older wage-earners. The second projection fits much better. We can call this equation a **quadratic projection** since the function is quadratic in *experience*.

Invertibility and Identification

The linear projection coefficient $\boldsymbol{\beta} = (\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y)$ exists and is unique as long as the $k \times k$ matrix $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ is invertible. The matrix $\boldsymbol{Q}_{\boldsymbol{xx}}$ is sometimes called the **design matrix**, as in experimental settings the researcher is able to control $\boldsymbol{Q}_{\boldsymbol{xx}}$ by manipulating the distribution of the regressors $\boldsymbol{x}$.

Observe that for any non-zero $\boldsymbol{\alpha} \in \mathbb{R}^k$,

$$\boldsymbol{\alpha}'\boldsymbol{Q}_{\boldsymbol{xx}}\boldsymbol{\alpha} = \mathbb{E}\left(\boldsymbol{\alpha}'\boldsymbol{x}\boldsymbol{x}'\boldsymbol{\alpha}\right) = \mathbb{E}\left(\boldsymbol{\alpha}'\boldsymbol{x}\right)^2 \geq 0$$

so $\boldsymbol{Q}_{\boldsymbol{xx}}$ by construction is positive semi-definite. The assumption that it is positive definite means that this is a strict inequality, $\mathbb{E}(\boldsymbol{\alpha}'\boldsymbol{x})^2 > 0$. Equivalently, there cannot exist a non-zero vector $\boldsymbol{\alpha}$ such that $\boldsymbol{\alpha}'\boldsymbol{x} = 0$ identically. This occurs when redundant variables are included in $\boldsymbol{x}$. Positive semi-definite matrices are invertible if and only if they are positive definite. When $\boldsymbol{Q}_{\boldsymbol{xx}}$ is invertible then $\boldsymbol{\beta} = (\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-1}\mathbb{E}(\boldsymbol{x}y)$ exists and is uniquely defined. In other words, in order for $\boldsymbol{\beta}$ to be uniquely defined, we must exclude the degenerate situation of redundant variables.

Theorem 2.18.1 shows that the linear projection coefficient $\boldsymbol{\beta}$ is **identified** (uniquely determined) under Assumption 2.18.1. The key is invertibility of $\boldsymbol{Q}_{\boldsymbol{xx}}$. Otherwise, there is no unique solution to the equation

$$\boldsymbol{Q}_{\boldsymbol{xx}}\boldsymbol{\beta} = \boldsymbol{Q}_{\boldsymbol{x}y}. \tag{2.33}$$

When $\boldsymbol{Q}_{\boldsymbol{xx}}$ is not invertible there are multiple solutions to (2.33), all of which yield an equivalent best linear predictor $\boldsymbol{x}'\boldsymbol{\beta}$. In this case the coefficient $\boldsymbol{\beta}$ is **not identified** as it does not have a unique value. Even so, the best linear predictor $\boldsymbol{x}'\boldsymbol{\beta}$ still identified. One solution is to set

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-}\mathbb{E}(\boldsymbol{x}y)$$

where $\boldsymbol{A}^{-}$ denotes the generalized inverse of $\boldsymbol{A}$ (see Appendix A.6).

2.19 Linear Predictor Error Variance

As in the CEF model, we define the error variance as

$$\sigma^2 = \mathbb{E}\left(e^2\right).$$

Setting $Q_{yy} = \mathbb{E}\left(y^2\right)$ and $\boldsymbol{Q}_{\boldsymbol{y}\boldsymbol{x}} = \mathbb{E}(y\boldsymbol{x}')$ we can write σ^2 as

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left(\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2\right) \\
&= \mathbb{E}\left(y^2\right) - 2\mathbb{E}\left(y\boldsymbol{x}'\right)\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta} \\
&= Q_{yy} - 2\boldsymbol{Q}_{\boldsymbol{y}\boldsymbol{x}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{x}y} + \boldsymbol{Q}_{\boldsymbol{y}\boldsymbol{x}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xx}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{x}y} \\
&= Q_{yy} - \boldsymbol{Q}_{\boldsymbol{y}\boldsymbol{x}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{x}y} \\
&\overset{def}{=} Q_{yy\cdot\boldsymbol{x}}.
\end{aligned} \tag{2.34}$$

One useful feature of this formula is that it shows that $Q_{yy\cdot\boldsymbol{x}} = Q_{yy} - \boldsymbol{Q}_{\boldsymbol{y}\boldsymbol{x}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{x}y}$ equals the variance of the error from the linear projection of y on $\boldsymbol{x}$.

2.20 Regression Coefficients

Sometimes it is useful to separate the constant from the other regressors, and write the linear projection equation in the format

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \alpha + e \tag{2.35}$$

where α is the intercept and $\boldsymbol{x}$ does not contain a constant.

Taking expectations of this equation, we find

$$\mathbb{E}(y) = \mathbb{E}\left(\boldsymbol{x}'\boldsymbol{\beta}\right) + \mathbb{E}(\alpha) + \mathbb{E}(e)$$

or

$$\mu_y = \boldsymbol{\mu}_x'\boldsymbol{\beta} + \alpha$$

where $\mu_y = \mathbb{E}(y)$ and $\boldsymbol{\mu}_{\boldsymbol{x}} = \mathbb{E}(\boldsymbol{x})$, since $\mathbb{E}(e) = 0$ from (2.30). (While $\boldsymbol{x}$ does not contain a constant, the equation does so (2.30) still applies.) Rearranging, we find

$$\alpha = \mu_y - \boldsymbol{\mu}_{\boldsymbol{x}}'\boldsymbol{\beta}.$$

Subtracting this equation from (2.35) we find

$$y - \mu_y = (\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})'\boldsymbol{\beta} + e, \tag{2.36}$$

a linear equation between the centered variables $y - \mu_y$ and $\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}$. (They are centered at their means, so are mean-zero random variables.) Because $\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}}$ is uncorrelated with e, (2.36) is also a linear projection, thus by the formula for the linear projection model,

$$\begin{aligned}\boldsymbol{\beta} &= \left(\mathbb{E}\left((\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})'\right)\right)^{-1}\mathbb{E}\left((\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{x}})(y - \mu_y)\right) \\ &= \operatorname{var}(\boldsymbol{x})^{-1}\operatorname{cov}(\boldsymbol{x}, y)\end{aligned}$$

a function only of the covariances[11] of $\boldsymbol{x}$ and y.

Theorem 2.20.1 *In the linear projection model*

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \alpha + e,$$

then

$$\alpha = \mu_y - \boldsymbol{\mu}_{\boldsymbol{x}}'\boldsymbol{\beta} \tag{2.37}$$

and

$$\boldsymbol{\beta} = \operatorname{var}(\boldsymbol{x})^{-1}\operatorname{cov}(\boldsymbol{x}, y). \tag{2.38}$$

2.21 Regression Sub-Vectors

Let the regressors be partitioned as

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{pmatrix}. \tag{2.39}$$

[11] The **covariance matrix** between vectors $\boldsymbol{x}$ and $\boldsymbol{z}$ is $\operatorname{cov}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}\left((\boldsymbol{x} - \mathbb{E}\boldsymbol{x})(\boldsymbol{z} - \mathbb{E}\boldsymbol{z})'\right)$. The (co)variance matrix of the vector $\boldsymbol{x}$ is $\operatorname{var}(\boldsymbol{x}) = \operatorname{cov}(\boldsymbol{x}, \boldsymbol{x}) = \mathbb{E}\left((\boldsymbol{x} - \mathbb{E}\boldsymbol{x})(\boldsymbol{x} - \mathbb{E}\boldsymbol{x})'\right)$.

We can write the projection of y on $\boldsymbol{x}$ as

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ &= \boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}_2'\boldsymbol{\beta}_2 + e \\ \mathbb{E}(\boldsymbol{x}e) &= \mathbf{0}. \end{aligned} \tag{2.40}$$

In this section we derive formula for the sub-vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

Partition $\boldsymbol{Q}_{\boldsymbol{xx}}$ conformably with $\boldsymbol{x}$

$$\boldsymbol{Q}_{\boldsymbol{xx}} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1') & \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_2') \\ \mathbb{E}(\boldsymbol{x}_2\boldsymbol{x}_1') & \mathbb{E}(\boldsymbol{x}_2\boldsymbol{x}_2') \end{bmatrix}$$

and similarly $\boldsymbol{Q}_{\boldsymbol{xy}}$

$$\boldsymbol{Q}_{\boldsymbol{xy}} = \begin{bmatrix} \boldsymbol{Q}_{1y} \\ \boldsymbol{Q}_{2y} \end{bmatrix} = \begin{bmatrix} \mathbb{E}(\boldsymbol{x}_1 y) \\ \mathbb{E}(\boldsymbol{x}_2 y) \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.4)

$$\boldsymbol{Q}_{\boldsymbol{xx}}^{-1} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}^{-1} \overset{def}{=} \begin{bmatrix} \boldsymbol{Q}^{11} & \boldsymbol{Q}^{12} \\ \boldsymbol{Q}^{21} & \boldsymbol{Q}^{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_{11\cdot2}^{-1} & -\boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot1}^{-1} \end{bmatrix}. \tag{2.41}$$

where $\boldsymbol{Q}_{11\cdot2} \overset{def}{=} \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and $\boldsymbol{Q}_{22\cdot1} \overset{def}{=} \boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$. Thus

$$\begin{aligned} \boldsymbol{\beta} &= \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \\ &= \begin{bmatrix} \boldsymbol{Q}_{11\cdot2}^{-1} & -\boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot1}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_{1y} \\ \boldsymbol{Q}_{2y} \end{bmatrix} \\ &= \begin{pmatrix} \boldsymbol{Q}_{11\cdot2}^{-1}\left(\boldsymbol{Q}_{1y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2y}\right) \\ \boldsymbol{Q}_{22\cdot1}^{-1}\left(\boldsymbol{Q}_{2y} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{1y}\right) \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{1y\cdot2} \\ \boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{2y\cdot1} \end{pmatrix}. \end{aligned}$$

We have shown that

$$\begin{aligned} \boldsymbol{\beta}_1 &= \boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{1y\cdot2} \\ \boldsymbol{\beta}_2 &= \boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{2y\cdot1}. \end{aligned}$$

2.22 Coefficient Decomposition

In the previous section we derived formulae for the coefficient sub-vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. We now use these formulae to give a useful interpretation of the coefficients in terms of an iterated projection.

Take equation (2.40) for the case $\dim(x_1) = 1$ so that $\beta_1 \in \mathbb{R}$.

$$y = x_1\beta_1 + \boldsymbol{x}_2'\boldsymbol{\beta}_2 + e. \tag{2.42}$$

Now consider the projection of x_1 on $\boldsymbol{x}_2$:

$$\begin{aligned} x_1 &= \boldsymbol{x}_2'\boldsymbol{\gamma}_2 + u_1 \\ \mathbb{E}(\boldsymbol{x}_2 u_1) &= \mathbf{0}. \end{aligned}$$

From (2.24) and (2.34), $\boldsymbol{\gamma}_2 = \boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and $\mathbb{E}u_1^2 = \boldsymbol{Q}_{11\cdot2} = \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$. We can also calculate that

$$\mathbb{E}(u_1 y) = \mathbb{E}\left(\left(x_1 - \boldsymbol{\gamma}_2'\boldsymbol{x}_2\right)y\right) = \mathbb{E}(x_1 y) - \boldsymbol{\gamma}_2'\mathbb{E}(\boldsymbol{x}_2 y) = \boldsymbol{Q}_{1y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2y} = \boldsymbol{Q}_{1y\cdot2}.$$

We have found that

$$\beta_1 = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1y\cdot 2} = \frac{\mathbb{E}\left(u_1 y\right)}{\mathbb{E}\left(u_1^2\right)}$$

the coefficient from the simple regression of y on u_1.

What this means is that in the multivariate projection equation (2.42), the coefficient β_1 equals the projection coefficient from a regression of y on u_1, the error from a projection of x_1 on the other regressors $\boldsymbol{x}_2$. The error u_1 can be thought of as the component of x_1 which is not linearly explained by the other regressors. Thus the coefficient β_1 equals the linear effect of x_1 on y, after stripping out the effects of the other variables.

There was nothing special in the choice of the variable x_1. This derivation applies symmetrically to all coefficients in a linear projection. Each coefficient equals the simple regression of y on the error from a projection of that regressor on all the other regressors. Each coefficient equals the linear effect of that variable on y, after linearly controlling for all the other regressors.

2.23 Omitted Variable Bias

Again, let the regressors be partitioned as in (2.39). Consider the projection of y on $\boldsymbol{x}_1$ only. Perhaps this is done because the variables $\boldsymbol{x}_2$ are not observed. This is the equation

$$\begin{aligned} y &= \boldsymbol{x}_1'\boldsymbol{\gamma}_1 + u \\ \mathbb{E}\left(\boldsymbol{x}_1 u\right) &= \mathbf{0}. \end{aligned} \tag{2.43}$$

Notice that we have written the coefficient on $\boldsymbol{x}_1$ as $\boldsymbol{\gamma}_1$ rather than $\boldsymbol{\beta}_1$ and the error as u rather than e. This is because (2.43) is different than (2.40). Goldberger (1991) introduced the catchy labels **long regression** for (2.40) and **short regression** for (2.43) to emphasize the distinction.

Typically, $\boldsymbol{\beta}_1 \neq \boldsymbol{\gamma}_1$, except in special cases. To see this, we calculate

$$\begin{aligned} \boldsymbol{\gamma}_1 &= \left(\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_1 y\right) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_1\left(\boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}_2'\boldsymbol{\beta}_2 + e\right)\right) \\ &= \boldsymbol{\beta}_1 + \left(\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_2'\right)\boldsymbol{\beta}_2 \\ &= \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2 \end{aligned}$$

where $\boldsymbol{\Gamma}_{12} = \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$ is the coefficient matrix from a projection of $\boldsymbol{x}_2$ on $\boldsymbol{x}_1$, where we use the notation from Section 2.21.

Observe that $\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_1$ unless $\boldsymbol{\Gamma}_{12} = \mathbf{0}$ or $\boldsymbol{\beta}_2 = \mathbf{0}$. Thus the short and long regressions have different coefficients on $\boldsymbol{x}_1$. They are the same only under one of two conditions. First, if the projection of $\boldsymbol{x}_2$ on $\boldsymbol{x}_1$ yields a set of zero coefficients (they are uncorrelated), or second, if the coefficient on $\boldsymbol{x}_2$ in (2.40) is zero. In general, the coefficient in (2.43) is $\boldsymbol{\gamma}_1$ rather than $\boldsymbol{\beta}_1$. The difference $\boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2$ between $\boldsymbol{\gamma}_1$ and $\boldsymbol{\beta}_1$ is known as **omitted variable bias**. It is the consequence of omission of a relevant correlated variable.

To avoid omitted variables bias the standard advice is to include all potentially relevant variables in estimated models. By construction, the general model will be free of such bias. Unfortunately in many cases it is not feasible to completely follow this advice as many desired variables are not observed. In this case, the possibility of omitted variables bias should be acknowledged and discussed in the course of an empirical investigation.

For example, suppose y is log wages, x_1 is education, and x_2 is intellectual ability. It seems reasonable to suppose that education and intellectual ability are positively correlated (highly able individuals attain higher levels of education) which means $\boldsymbol{\Gamma}_{12} > 0$. It also seems reasonable to suppose that conditional on education, individuals with higher intelligence will earn higher wages on average, so that $\beta_2 > 0$. This implies that $\boldsymbol{\Gamma}_{12}\beta_2 > 0$ and $\gamma_1 = \beta_1 + \boldsymbol{\Gamma}_{12}\beta_2 > \beta_1$. Therefore,

it seems reasonable to expect that in a regression of wages on education with ability omitted, the coefficient on education is higher than in a regression where ability is included. In other words, in this context the omitted variable biases the regression coefficient upwards. It is possible, for example, that $\beta_1 = 0$ so that education has no direct effect on wages yet $\gamma_1 = \boldsymbol{\Gamma}_{12}\beta_2 > 0$ meaning that the regression coefficient on education alone is positive, but is a consequence of the unmodeled correlation between education and intellectual ability.

Unfortunately the above simple characterization of omitted variable bias does not immediately carry over to more complicated settings, as discovered by Luca, Magnus, and Peracchi (2017). For example, suppose we compare three nested projections

$$\begin{aligned} y &= \boldsymbol{x}_1'\boldsymbol{\gamma}_1 + u_1 \\ y &= \boldsymbol{x}_1'\boldsymbol{\delta}_1 + \boldsymbol{x}_2'\boldsymbol{\delta}_2 + u_2 \\ y &= \boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}_2'\boldsymbol{\beta}_2 + \boldsymbol{x}_3'\boldsymbol{\beta}_3 + e. \end{aligned}$$

We can call them the short, medium, and long regressions. Suppose that the parameter of interest is $\boldsymbol{\beta}_1$ in the long regression. We are interested in the consequences of omitting $\boldsymbol{x}_3$ when estimating the medium regression, and of omitting both $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ when estimating the short regression. In particular we are interested in the question: Is it better to estimate the short or medium regression, given that both omit $\boldsymbol{x}_3$? Intuition suggests that the medium regression should be "less biased" but it is worth investigating in greater detail. By similar calculations to those above, we find that

$$\begin{aligned} \boldsymbol{\gamma}_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2 + \boldsymbol{\Gamma}_{13}\boldsymbol{\beta}_3 \\ \delta_1 &= \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{13\cdot 2}\boldsymbol{\beta}_3 \end{aligned}$$

where $\boldsymbol{\Gamma}_{13\cdot 2} = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{13\cdot 2}$ using the notation from Section 2.21.

We see that the bias in the short regression coefficient is $\boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2 + \boldsymbol{\Gamma}_{13}\boldsymbol{\beta}_3$ which depends on both $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$, while that for the medium regression coefficient is $\boldsymbol{\Gamma}_{13\cdot 2}\boldsymbol{\beta}_3$ which only depends on $\boldsymbol{\beta}_3$. So the bias for the medium regression is less complicated, and intuitively seems more likely to be smaller than that of the short regression. However it is impossible to strictly rank the two. It is quite possible that $\boldsymbol{\gamma}_1$ is less biased than $\boldsymbol{\delta}_1$. Thus as a general rule it is strictly impossible to state that estimation of the medium regression will be less biased than estimation of the short regression.

2.24 Best Linear Approximation

There are alternative ways we could construct a linear approximation $\boldsymbol{x}'\boldsymbol{\beta}$ to the conditional mean $m(\boldsymbol{x})$. In this section we show that one alternative approach turns out to yield the same answer as the best linear predictor.

We start by defining the mean-square approximation error of $\boldsymbol{x}'\boldsymbol{\beta}$ to $m(\boldsymbol{x})$ as the expected squared difference between $\boldsymbol{x}'\boldsymbol{\beta}$ and the conditional mean $m(\boldsymbol{x})$

$$d(\boldsymbol{\beta}) = \mathbb{E}\left(\left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2\right). \tag{2.44}$$

The function $d(\boldsymbol{\beta})$ is a measure of the deviation of $\boldsymbol{x}'\boldsymbol{\beta}$ from $m(\boldsymbol{x})$. If the two functions are identical then $d(\boldsymbol{\beta}) = 0$, otherwise $d(\boldsymbol{\beta}) > 0$. We can also view the mean-square difference $d(\boldsymbol{\beta})$ as a density-weighted average of the function $\left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2$, since

$$d(\boldsymbol{\beta}) = \int_{\mathbb{R}^k} \left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2 f_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}$$

where $f_{\boldsymbol{x}}(\boldsymbol{x})$ is the marginal density of $\boldsymbol{x}$.

We can then define the best linear approximation to the conditional $m(\boldsymbol{x})$ as the function $\boldsymbol{x}'\boldsymbol{\beta}$ obtained by selecting $\boldsymbol{\beta}$ to minimize $d(\boldsymbol{\beta})$:

$$\boldsymbol{\beta} = \underset{\boldsymbol{b}\in\mathbb{R}^k}{\operatorname{argmin}}\, d(\boldsymbol{b}). \tag{2.45}$$

Similar to the best linear predictor we are measuring accuracy by expected squared error. The difference is that the best linear predictor (2.21) selects $\boldsymbol{\beta}$ to minimize the expected squared prediction error, while the best linear approximation (2.45) selects $\boldsymbol{\beta}$ to minimize the expected squared approximation error.

Despite the different definitions, it turns out that the best linear predictor and the best linear approximation are identical. By the same steps as in (2.18) plus an application of conditional expectations we can find that

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}m(\boldsymbol{x})\right) \tag{2.46}$$

$$= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right) \tag{2.47}$$

(see Exercise 2.19). Thus (2.45) equals (2.21). We conclude that the definition (2.45) can be viewed as an alternative motivation for the linear projection coefficient.

2.25 Regression to the Mean

The term **regression** originated in an influential paper by Francis Galton (1886), where he examined the joint distribution of the stature (height) of parents and children. Effectively, he was estimating the conditional mean of children's height given their parent's height. Galton discovered that this conditional mean was approximately linear with a slope of 2/3. This implies that *on average* a child's height is more mediocre (average) than his or her parent's height. Galton called this phenomenon **regression to the mean**, and the label **regression** has stuck to this day to describe most conditional relationships.

One of Galton's fundamental insights was to recognize that if the marginal distributions of y and x are the same (e.g. the heights of children and parents in a stable environment) then the regression slope in a linear projection is always less than one.

To be more precise, take the simple linear projection

$$y = x\beta + \alpha + e \tag{2.48}$$

where y equals the height of the child and x equals the height of the parent. Assume that y and x have the same mean, so that $\mu_y = \mu_x = \mu$. Then from (2.37)

$$\alpha = (1-\beta)\,\mu$$

so we can write the linear projection (2.48) as

$$\mathcal{P}\left(y \mid x\right) = (1-\beta)\,\mu + x\beta.$$

This shows that the projected height of the child is a weighted average of the population average height μ and the parent's height x, with the weight equal to the regression slope β. When the height distribution is stable across generations, so that $\mathrm{var}(y) = \mathrm{var}(x)$, then this slope is the simple correlation of y and x. Using (2.38)

$$\beta = \frac{\mathrm{cov}\left(x, y\right)}{\mathrm{var}(x)} = \mathrm{corr}(x, y).$$

By the properties of correlation (e.g. equation (**??**) in the Appendix), $-1 \leq \mathrm{corr}(x, y) \leq 1$, with $\mathrm{corr}(x, y) = 1$ only in the degenerate case $y = x$. Thus if we exclude degeneracy, β is strictly less than 1.

This means that on average a child's height is more mediocre (closer to the population average) than the parent's.

Sir Francis Galton

Sir Francis Galton (1822-1911) of England was one of the leading figures in late 19th century statistics. In addition to inventing the concept of regression, he is credited with introducing the concepts of correlation, the standard deviation, and the bivariate normal distribution. His work on heredity made a significant intellectual advance by examing the joint distributions of observables, allowing the application of the tools of mathematical statistics to the social sciences.

A common error – known as the **regression fallacy** – is to infer from $\beta < 1$ that the population is **converging**, meaning that its variance is declining towards zero. This is a fallacy because we derived the implication $\beta < 1$ under the assumption of constant means and variances. So certainly $\beta < 1$ does not imply that the variance y is less than than the variance of x.

Another way of seeing this is to examine the conditions for convergence in the context of equation (2.48). Since x and e are uncorrelated, it follows that

$$\operatorname{var}(y) = \beta^2 \operatorname{var}(x) + \operatorname{var}(e).$$

Then $\operatorname{var}(y) < \operatorname{var}(x)$ if and only if

$$\beta^2 < 1 - \frac{\operatorname{var}(e)}{\operatorname{var}(x)}$$

which is not implied by the simple condition $|\beta| < 1$.

The regression fallacy arises in related empirical situations. Suppose you sort families into groups by the heights of the parents, and then plot the average heights of each subsequent generation over time. If the population is stable, the regression property implies that the plots lines will converge – children's height will be more average than their parents. The regression fallacy is to incorrectly conclude that the population is converging. A message to be learned from this example is that such plots are misleading for inferences about convergence.

The regression fallacy is subtle. It is easy for intelligent economists to succumb to its temptation. A famous example is *The Triumph of Mediocrity in Business* by Horace Secrist, published in 1933. In this book, Secrist carefully and with great detail documented that in a sample of department stores over 1920-1930, when he divided the stores into groups based on 1920-1921 profits, and plotted the average profits of these groups for the subsequent 10 years, he found clear and persuasive evidence for convergence "toward mediocrity". Of course, there was no discovery – regression to the mean is a necessary feature of stable distributions.

2.26 Reverse Regression

Galton noticed another interesting feature of the bivariate distribution. There is nothing special about a regression of y on x. We can also regress x on y. (In his heredity example this is the best linear predictor of the height of parents given the height of their children.) This regression takes the form

$$x = y\beta^* + \alpha^* + e^*. \tag{2.49}$$

This is sometimes called the **reverse regression**. In this equation, the coefficients α^*, β^* and error e^* are defined by linear projection. In a stable population we find that

$$\beta^* = \operatorname{corr}(x, y) = \beta$$

$$\alpha^* = (1 - \beta)\,\mu = \alpha$$

which are exactly the same as in the projection of y on x! The intercept and slope have exactly the same values in the forward and reverse projections!

While this algebraic discovery is quite simple, it is counter-intuitive. Instead, a common yet mistaken guess for the form of the reverse regression is to take the equation (2.48), divide through by β and rewrite to find the equation

$$x = y\frac{1}{\beta} - \frac{\alpha}{\beta} - \frac{1}{\beta}e \tag{2.50}$$

suggesting that the projection of x on y should have a slope coefficient of $1/\beta$ instead of β, and intercept of $-\alpha/\beta$ rather than α. What went wrong? Equation (2.50) is perfectly valid, because it is a simple manipulation of the valid equation (2.48). The trouble is that (2.50) is neither a CEF nor a linear projection. Inverting a projection (or CEF) does not yield a projection (or CEF). Instead, (2.49) is a valid projection, not (2.50).

In any event, Galton's finding was that when the variables are standardized, the slope in both projections (y on x, and x and y) equals the correlation, and both equations exhibit regression to the mean. It is not a causal relation, but a natural feature of all joint distributions.

2.27 Limitations of the Best Linear Projection

Let's compare the linear projection and linear CEF models.

From Theorem 2.8.1.4 we know that the CEF error has the property $\mathbb{E}(\boldsymbol{x}e) = \mathbf{0}$. Thus a linear CEF is the best linear projection. However, the converse is not true as the projection error does not necessarily satisfy $\mathbb{E}(e \mid \boldsymbol{x}) = 0$. Furthermore, the linear projection may be a poor approximation to the CEF.

To see these points in a simple example, suppose that the true process is $y = x + x^2$ with $x \sim \mathrm{N}(0, 1)$. In this case the true CEF is $m(x) = x + x^2$ and there is no error. Now consider the linear projection of y on x and a constant, namely the model $y = \beta x + \alpha + u$. Since $x \sim \mathrm{N}(0, 1)$ then x and x^2 are uncorrelated and the linear projection takes the form $\mathcal{P}(y \mid x) = x + 1$. This is quite different from the true CEF $m(x) = x + x^2$. The projection error equals $e = x^2 - 1$, which is a deterministic function of x, yet is uncorrelated with x. We see in this example that a projection error need not be a CEF error, and a linear projection can be a poor approximation to the CEF.

Another defect of linear projection is that it is sensitive to the marginal distribution of the regressors when the conditional mean is non-linear. We illustrate the issue in Figure 2.10 for a constructed[12] joint distribution of y and x. The solid line is the non-linear CEF of y given x. The data are divided in two groups – Group 1 and Group 2 – which have different marginal distributions for the regressor x, and Group 1 has a lower mean value of x than Group 2. The separate linear projections of y on x for these two groups are displayed in the Figure by the dashed lines. These two projections are distinct approximations to the CEF. A defect with linear projection is that it leads to the incorrect conclusion that the effect of x on y is different for individuals in the two groups. This conclusion is incorrect because in fact there is no difference in the conditional mean function. The apparent difference is a by-product of a linear approximation to a nonlinear mean, combined with different marginal distributions for the conditioning variables.

2.28 Random Coefficient Model

A model which is notationally similar to but conceptually distinct from the linear CEF model is the linear random coefficient model. It takes the form

$$y = \boldsymbol{x}'\boldsymbol{\eta}$$

[12] The x in Group 1 are $\mathrm{N}(2, 1)$ and those in Group 2 are $\mathrm{N}(4, 1)$, and the conditional distribution of y given x is $\mathrm{N}(m(x), 1)$ where $m(x) = 2x - x^2/6$.

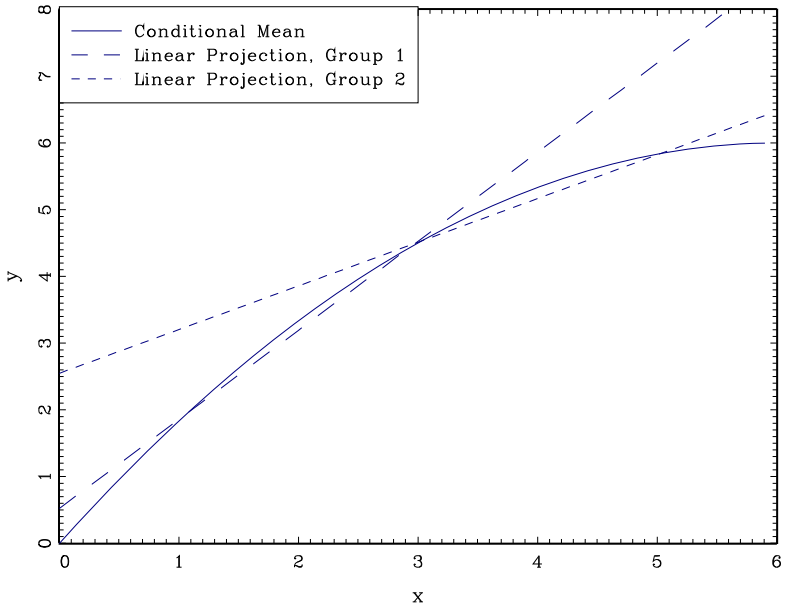


Figure 2.10: Conditional Mean and Two Linear Projections

where the individual-specific coefficient $\boldsymbol{\eta}$ is random and independent of $\boldsymbol{x}$. For example, if $\boldsymbol{x}$ is years of schooling and y is log wages, then $\boldsymbol{\eta}$ is the individual-specific returns to schooling. If a person obtains an extra year of schooling, $\boldsymbol{\eta}$ is the actual change in their wage. The random coefficient model allows the returns to schooling to vary in the population. Some individuals might have a high return to education (a high $\boldsymbol{\eta}$) and others a low return, possibly 0, or even negative.

In the linear CEF model the regressor coefficient equals the regression derivative – the change in the conditional mean due to a change in the regressors, $\boldsymbol{\beta} = \boldsymbol{\nabla} m(\boldsymbol{x})$. This is not the effect on a given individual, it is the effect on the population average. In contrast, in the random coefficient model, the random vector $\boldsymbol{\eta} = \boldsymbol{\nabla}(\boldsymbol{x}'\boldsymbol{\eta})$ is the true causal effect – the change in the response variable y itself due to a change in the regressors.

It is interesting, however, to discover that the linear random coefficient model implies a linear CEF. To see this, let $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ denote the mean and covariance matrix of $\boldsymbol{\eta}$:

$$\begin{aligned}\boldsymbol{\beta} &= \mathbb{E}(\boldsymbol{\eta}) \\ \boldsymbol{\Sigma} &= \operatorname{var}(\boldsymbol{\eta})\end{aligned}$$

and then decompose the random coefficient as

$$\boldsymbol{\eta} = \boldsymbol{\beta} + \boldsymbol{u}$$

where $\boldsymbol{u}$ is distributed independently of $\boldsymbol{x}$ with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Then we can write

$$\mathbb{E}(y \mid \boldsymbol{x}) = \boldsymbol{x}'\mathbb{E}(\boldsymbol{\eta} \mid \boldsymbol{x}) = \boldsymbol{x}'\mathbb{E}(\boldsymbol{\eta}) = \boldsymbol{x}'\boldsymbol{\beta}$$

so the CEF is linear in $\boldsymbol{x}$**,** and the coefficients $\boldsymbol{\beta}$ equal the mean of the random coefficient $\boldsymbol{\eta}$.

We can thus write the equation as a linear CEF

$$y = \boldsymbol{x}'\boldsymbol{\beta} + e \tag{2.51}$$

where $e = \boldsymbol{x}'\boldsymbol{u}$ and $\boldsymbol{u} = \boldsymbol{\eta} - \boldsymbol{\beta}$. The error is conditionally mean zero:

$$\mathbb{E}(e \mid \boldsymbol{x}) = 0.$$

Furthermore

$$\begin{aligned}\operatorname{var}(e \mid \boldsymbol{x}) &= \boldsymbol{x}' \operatorname{var}(\boldsymbol{\eta}) \boldsymbol{x} \\ &= \boldsymbol{x}' \boldsymbol{\Sigma} \boldsymbol{x}\end{aligned}$$

so the error is conditionally heteroskedastic with its variance a quadratic function of $\boldsymbol{x}$.

Theorem 2.28.1 *In the linear random coefficient model $y = \boldsymbol{x}'\boldsymbol{\eta}$ with $\boldsymbol{\eta}$ independent of $\boldsymbol{x}$, $\mathbb{E}\|\boldsymbol{x}\|^2 < \infty$, and $\mathbb{E}\|\boldsymbol{\eta}\|^2 < \infty$, then*

$$\begin{aligned}\mathbb{E}(y \mid \boldsymbol{x}) &= \boldsymbol{x}'\boldsymbol{\beta} \\ \operatorname{var}(y \mid \boldsymbol{x}) &= \boldsymbol{x}'\boldsymbol{\Sigma}\boldsymbol{x}\end{aligned}$$

where $\boldsymbol{\beta} = \mathbb{E}(\boldsymbol{\eta})$ and $\boldsymbol{\Sigma} = \operatorname{var}(\boldsymbol{\eta})$.

2.29 Causal Effects

So far we have avoided the concept of causality, yet often the underlying goal of an econometric analysis is to uncover a causal relationship between variables. It is often of great interest to understand the causes and effects of decisions, actions, and policies. For example, we may be interested in the effect of class sizes on test scores, police expenditures on crime rates, climate change on economic activity, years of schooling on wages, institutional structure on growth, the effectiveness of rewards on behavior, the consequences of medical procedures for health outcomes, or any variety of possible causal relationships. In each case, the goal is to understand what is the actual effect on the outcome y due to a change in the input x. We are not just interested in the conditional mean or linear projection, we would like to know the actual change.

Two inherent barriers are that the causal effect is typically specific to an individual and that it is unobserved.

Consider the effect of schooling on wages. The causal effect is the actual difference a person would receive in wages if we could change their level of education *holding all else constant*. This is specific to each individual as their employment outcomes in these two distinct situations is individual. The causal effect is unobserved because the most we can observe is their actual level of education and their actual wage, but not the counterfactual wage if their education had been different.

To be even more specific, suppose that there are two individuals, Jennifer and George, and both have the possibility of being high-school graduates or college graduates, but both would have received different wages given their choices. For example, suppose that Jennifer would have earned \$10 an hour as a high-school graduate and \$20 an hour as a college graduate while George would have earned \$8 as a high-school graduate and \$12 as a college graduate. In this example the causal effect of schooling is \$10 a hour for Jennifer and \$4 an hour for George. The causal effects are specific to the individual and neither causal effect is observed.

A variable x_1 can be said to have a causal effect on the response variable y if the latter changes when all other inputs are held constant. To make this precise we need a mathematical formulation. We can write a full model for the response variable y as

$$y = h(x_1, \boldsymbol{x}_2, \boldsymbol{u}) \tag{2.52}$$

where x_1 and $\boldsymbol{x}_2$ are the observed variables, $\boldsymbol{u}$ is an $\ell \times 1$ unobserved random factor, and h is a functional relationship. This framework, called the **potential outcomes** framework, includes as

a special case the random coefficient model (2.28) studied earlier. We define the causal effect of x_1 within this model as the change in y due to a change in x_1 holding the other variables $\boldsymbol{x}_2$ and $\boldsymbol{u}$ constant.

Definition 2.29.1 *In the model (2.52) the* ***causal effect*** *of x_1 on y is*

$$C(x_1, \boldsymbol{x}_2, \boldsymbol{u}) = \boldsymbol{\nabla}_1 h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right), \tag{2.53}$$

the change in y due to a change in x_1, holding $\boldsymbol{x}_2$ and $\boldsymbol{u}$ constant.

To understand this concept, imagine taking a single individual. As far as our structural model is concerned, this person is described by their observables x_1 and $\boldsymbol{x}_2$ and their unobservables $\boldsymbol{u}$. In a wage regression the unobservables would include characteristics such as the person's abilities, skills, work ethic, interpersonal connections, and preferences. The causal effect of x_1 (say, education) is the change in the wage as x_1 changes, holding constant all other observables **and** unobservables.

It may be helpful to understand that (2.53) is a definition, and does not necessarily describe causality in a fundamental or experimental sense. Perhaps it would be more appropriate to label (2.53) as a **structural effect** (the effect within the structural model).

Sometimes it is useful to write this relationship as a potential outcome function

$$y(x_1) = h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right)$$

where the notation implies that $y(x_1)$ is holding $\boldsymbol{x}_2$ and $\boldsymbol{u}$ constant.

A popular example arises in the analysis of treatment effects with a binary regressor x_1. Let $x_1 = 1$ indicate treatment (e.g. a medical procedure) and $x_1 = 0$ indicate non-treatment. In this case $y(x_1)$ can be written

$$\begin{aligned} y(0) &= h\left(0, \boldsymbol{x}_2, \boldsymbol{u}\right) \\ y(1) &= h\left(1, \boldsymbol{x}_2, \boldsymbol{u}\right). \end{aligned}$$

In the literature on treatment effects, it is common to refer to $y(0)$ and $y(1)$ as the latent outcomes associated with non-treatment and treatment, respectively. That is, for a given individual, $y(0)$ is the health outcome if there is no treatment, and $y(1)$ is the health outcome if there is treatment. The causal effect of treatment for the individual is the change in their health outcome due to treatment – the change in y as we hold both $\boldsymbol{x}_2$ and $\boldsymbol{u}$ constant:

$$C\left(\boldsymbol{x}_2, \boldsymbol{u}\right) = y(1) - y(0).$$

This is random (a function of $\boldsymbol{x}_2$ and $\boldsymbol{u}$) as both potential outcomes $y(0)$ and $y(1)$ are different across individuals.

In a sample, we cannot observe both outcomes from the same individual, we only observe the realized value

$$y = \begin{cases} y(0) & \text{if} \quad x_1 = 0 \\ \\ y(1) & \text{if} \quad x_1 = 1. \end{cases}$$

As the causal effect varies across individuals and is not observable, it cannot be measured on the individual level. We therefore focus on aggregate causal effects, in particular what is known as the average causal effect.

Definition 2.29.2 *In the model (2.52) the* ***average causal effect*** *of x_1 on y conditional on $\boldsymbol{x}_2$ is*

$$\begin{aligned} ACE(x_1, \boldsymbol{x}_2) &= \mathbb{E}\left(C(x_1, \boldsymbol{x}_2, \boldsymbol{u}) \mid x_1, \boldsymbol{x}_2\right) \\ &= \int_{\mathbb{R}^\ell} \boldsymbol{\nabla}_1 h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) d\boldsymbol{u} \end{aligned} \tag{2.54}$$

where $f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2)$ is the conditional density of $\boldsymbol{u}$ given $x_1, \boldsymbol{x}_2$.

We can think of the average causal effect $ACE(x_1, \boldsymbol{x}_2)$ as the average effect in the general population. In our Jennifer & George schooling example given earlier, supposing that half of the population are Jennifer's and the other half George's, then the average causal effect of college is $(10+4)/2 = \$7$ an hour. This is not the individual causal effect, it is the average of the causal effect across all individuals in the population. Given data on only educational attainment and wages, the ACE of \$7 is the best we can hope to learn.

When we conduct a regression analysis (that is, consider the regression of observed wages on educational attainment) we might hope that the regression reveals the average causal effect. Technically, that the regression derivative (the coefficient on education) equals the ACE. Is this the case? In other words, what is the relationship between the average causal effect $ACE(x_1, \boldsymbol{x}_2)$ and the regression derivative $\boldsymbol{\nabla}_1 m\left(x_1, \boldsymbol{x}_2\right)$? Equation (2.52) implies that the CEF is

$$\begin{aligned} m(x_1, \boldsymbol{x}_2) &= \mathbb{E}\left(h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) \mid x_1, \boldsymbol{x}_2\right) \\ &= \int_{\mathbb{R}^\ell} h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) d\boldsymbol{u}, \end{aligned}$$

the average causal equation, averaged over the conditional distribution of the unobserved component $\boldsymbol{u}$.

Applying the marginal effect operator, the regression derivative is

$$\begin{aligned} \boldsymbol{\nabla}_1 m(x_1, \boldsymbol{x}_2) &= \int_{\mathbb{R}^\ell} \boldsymbol{\nabla}_1 h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) d\boldsymbol{u} \\ &\quad + \int_{\mathbb{R}^\ell} h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) \boldsymbol{\nabla}_1 f(\boldsymbol{u} | x_1, \boldsymbol{x}_2) d\boldsymbol{u} \\ &= ACE(x_1, \boldsymbol{x}_2) + \int_{\mathbb{R}^\ell} h\left(x_1, \boldsymbol{x}_2, \boldsymbol{u}\right) \boldsymbol{\nabla}_1 f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) d\boldsymbol{u}. \end{aligned} \tag{2.55}$$

Equation (2.55) shows that in general, the regression derivative does not equal the average causal effect. The difference is the second term on the right-hand-side of (2.55). The regression derivative and ACE equal in the special case when this term equals zero, which occurs when $\boldsymbol{\nabla}_1 f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) = 0$, that is, when the conditional density of $\boldsymbol{u}$ given $(x_1, \boldsymbol{x}_2)$ does not depend on x_1. When this condition holds then the regression derivative equals the ACE, which means that regression analysis can be interpreted causally, in the sense that it uncovers average causal effects.

The condition is sufficiently important that it has a special name in the treatment effects literature.

Definition 2.29.3 ***Conditional Independence Assumption (CIA).*** *Conditional on $\boldsymbol{x}_2$, the random variables x_1 and $\boldsymbol{u}$ are statistically independent.*

The CIA implies $f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) = f(\boldsymbol{u} \mid \boldsymbol{x}_2)$ does not depend on x_1, and thus $\boldsymbol{\nabla}_1 f(\boldsymbol{u} \mid x_1, \boldsymbol{x}_2) = 0$. Thus the CIA implies that $\boldsymbol{\nabla}_1 m(x_1, \boldsymbol{x}_2) = ACE(x_1, \boldsymbol{x}_2)$, the regression derivative equals the average causal effect.

Theorem 2.29.1 *In the structural model (2.52), the Conditional Independence Assumption implies*

$$\boldsymbol{\nabla}_1 m(x_1, \boldsymbol{x}_2) = ACE(x_1, \boldsymbol{x}_2)$$

the regression derivative equals the average causal effect for x_1 on y conditional on $\boldsymbol{x}_2$.

This is a fascinating result. It shows that whenever the unobservable is independent of the treatment variable (after conditioning on appropriate regressors) the regression derivative equals the average causal effect. In this case, the CEF has causal economic meaning, giving strong justification to estimation of the CEF. Our derivation also shows the critical role of the CIA. If CIA fails, then the equality of the regression derivative and ACE fails.

This theorem is quite general. It applies equally to the treatment-effects model where x_1 is binary or to more general settings where x_1 is continuous.

It is also helpful to understand that the CIA is weaker than full independence of $\boldsymbol{u}$ from the regressors $(x_1, \boldsymbol{x}_2)$. The CIA was introduced precisely as a minimal sufficient condition to obtain the desired result. Full independence implies the CIA and implies that each regression derivative equals that variable's average causal effect, but full independence is not necessary in order to causally interpret a subset of the regressors.

To illustrate, let's return to our education example involving a population with equal numbers of Jennifer's and George's. Recall that Jennifer earns \$10 as a high-school graduate and \$20 as a college graduate (and so has a causal effect of \$10) while George earns \$8 as a high-school graduate and \$12 as a college graduate (so has a causal effect of \$4). Given this information, the average causal effect of college is \$7, which is what we hope to learn from a regression analysis.

Now suppose that while in high school all students take an aptitude test, and if a student gets a high (H) score he or she goes to college with probability 3/4, and if a student gets a low (L) score he or she goes to college with probability 1/4. Suppose further that Jennifer's get an aptitude score of H with probability 3/4, while George's get a score of H with probability 1/4. Given this situation, 62.5% of Jennifer's will go to college[13], while 37.5% of George's will go to college[14].

An econometrician who randomly samples 32 individuals and collects data on educational attainment and wages will find the following wage distribution:

	\$8	\$10	\$12	\$20	Mean
High-School Graduate	10	6	0	0	\$8.75
College Graduate	0	0	6	10	\$17.00

Let *college* denote a dummy variable taking the value of 1 for a college graduate, otherwise 0. Thus the regression of wages on college attendance takes the form

$$\mathbb{E}(wage \mid college) = 8.25 college + 8.75.$$

The coefficient on the college dummy, \$8.25, is the regression derivative, and the implied wage effect of college attendance. But \$8.25 overstates the average causal effect of \$7. The reason is because

[13] $\Pr(College|Jennifer) = \Pr(College|H)\Pr(H|Jennifer) + \Pr(College|L)\Pr(L|Jennifer) = (3/4)^2 + (1/4)^2$

[14] $\Pr(College|George) = \Pr(College|H)\Pr(H|George) + \Pr(College|L)\Pr(L|George) = (3/4)(1/4) + (1/4)(3/4)$

the CIA fails. In this model the unobservable $\boldsymbol{u}$ is the individual's type (Jennifer or George) which is not independent of the regressor x_1 (education), since Jennifer is more likely to go to college than George. Since Jennifer's causal effect is higher than George's, the regression derivative overstates the ACE. The coefficient \$8.25 is not the average benefit of college attendance, rather it is the observed difference in realized wages in a population whose decision to attend college is correlated with their individual causal effect. At the risk of repeating myself, in this example, \$8.25 is the true regression derivative, it is the difference in average wages between those with a college education and those without. It is not, however, the average causal effect of college education in the population.

This does not mean that it is impossible to estimate the ACE. The key is conditioning on the appropriate variables. The CIA says that we need to find a variable x_2 such that conditional on x_2, $\boldsymbol{u}$ and x_1 (type and education) are independent. In this example a variable which will achieve this is the aptitude test score. The decision to attend college was based on the test score, not on an individual's type. Thus educational attainment and type are independent once we condition on the test score.

This also alters the ACE. Notice that Definition 2.29.2 is a function of x_2 (the test score). Among the students who receive a high test score, 3/4 are Jennifer's and 1/4 are George's. Thus the ACE for students with a score of H is $(3/4) \times 10 + (1/4) \times 4 = \$8.50$. Among the students who receive a low test score, 1/4 are Jennifer's and 3/4 are George's. Thus the ACE for students with a score of L is $(1/4) \times 10 + (3/4) \times 4 = \$5.50$. The ACE varies between these two observable groups (those with high test scores and those with low test scores). Again, we would hope to be able to learn the ACE from a regression analysis, this time from a regression of wages on education and test scores.

To see this in the wage distribution, suppose that the econometrician collects data on the aptitude test score as well as education and wages. Given a random sample of 32 individuals we would expect to find the following wage distribution:

	\$8	\$10	\$12	\$20	Mean
High-School Graduate + High Test Score	1	3	0	0	\$9.50
College Graduate + High Test Score	0	0	3	9	\$18.00
High-School Graduate + Low Test Score	9	3	0	0	\$8.50
College Graduate + Low Test Score	0	0	3	1	\$14.00

Define the dummy variable *highscore* which takes the value 1 for students who received a high test score, else zero. The regression of wages on college attendance and test scores (with interactions) takes the form

$$\mathbb{E}\left(wage \mid college, highscore\right) = 1.00 highscore + 5.50 college + 3.00 highscore \times college + 8.50.$$

The coefficient on *college*, \$5.50, is the regression derivative of college attendance for those with low test scores, and the sum of this coefficient with the interaction coefficient, \$8.50, is the regression derivative for college attendance for those with high test scores. These equal the average causal effect as calculated above. Furthermore, since 1/2 of the population achieves a high test score and 1/2 achieve a low test score, the measured average causal effect in the entire population is \$7, which precisely equals the true value.

In this example, by conditioning on the aptitude test score, the average causal effect of education on wages can be learned from a regression analysis. What this shows is that by conditioning on the proper variables, it may be possible to achieve the CIA, in which case regression analysis measures average causal effects.

2.30 Expectation: Mathematical Details*

We define the **mean** or **expectation** $\mathbb{E}(y)$ of a random variable y as follows. If y is discrete on the set $\{\tau_1, \tau_2, ...\}$ then

$$\mathbb{E}(y) = \sum_{j=1}^{\infty} \tau_j \Pr(y = \tau_j),$$

and if y is continuous with density f then

$$\mathbb{E}(y) = \int_{-\infty}^{\infty} y f(y) dy.$$

We can unify these definitions by writing the expectation as the Lebesgue integral with respect to the distribution function F

$$\mathbb{E}(y) = \int_{-\infty}^{\infty} y dF(y). \tag{2.56}$$

In the event that the integral (2.56) is not finite, separately evaluate the two integrals

$$I_1 = \int_{0}^{\infty} y dF(y) \tag{2.57}$$

$$I_2 = -\int_{-\infty}^{0} y dF(y). \tag{2.58}$$

If $I_1 = \infty$ and $I_2 < \infty$ then it is typical to define $\mathbb{E}(y) = \infty$. If $I_1 < \infty$ and $I_2 = \infty$ then we define $\mathbb{E}(y) = -\infty$. However, if both $I_1 = \infty$ and $I_2 = \infty$ then $\mathbb{E}(y)$ is undefined. If

$$\mathbb{E}|y| = \int_{-\infty}^{\infty} |y| \, dF(y) = I_1 + I_2 < \infty$$

then $\mathbb{E}(y)$ exists and is finite. In this case it is common to say that the mean $\mathbb{E}(y)$ is "well-defined".

More generally, y has a finite r^{th} moment if

$$\mathbb{E}|y|^r < \infty. \tag{2.59}$$

By Liapunov's Inequality (B.13), (2.59) implies $\mathbb{E}|y|^s < \infty$ for all $1 \leq s \leq r$. Thus, for example, if the fourth moment is finite then the first, second and third moments are also finite, and so is the 3.9^{th} moment.

It is common in econometric theory to assume that the variables, or certain transformations of the variables, have finite moments of a certain order. How should we interpret this assumption? How restrictive is it?

One way to visualize the importance is to consider the class of Pareto densities given by

$$f(y) = a y^{-a-1}, \qquad y > 1.$$

The parameter a of the Pareto distribution indexes the rate of decay of the tail of the density. Larger a means that the tail declines to zero more quickly. See Figure 2.11 below where we plot the Pareto density for $a = 1$ and $a = 2$. The parameter a also determines which moments are finite. We can calculate that

$$\mathbb{E}|y|^r = \begin{cases} a \int_1^{\infty} y^{r-a-1} dy = \dfrac{a}{a-r} & \text{if} \quad r < a \\ \\ \infty & \text{if} \quad r \geq a. \end{cases}$$

This shows that if y is Pareto distributed with parameter a, then the r^{th} moment of y is finite if and only if $r < a$. Higher a means higher finite moments. Equivalently, the faster the tail of the density declines to zero, the more moments are finite.

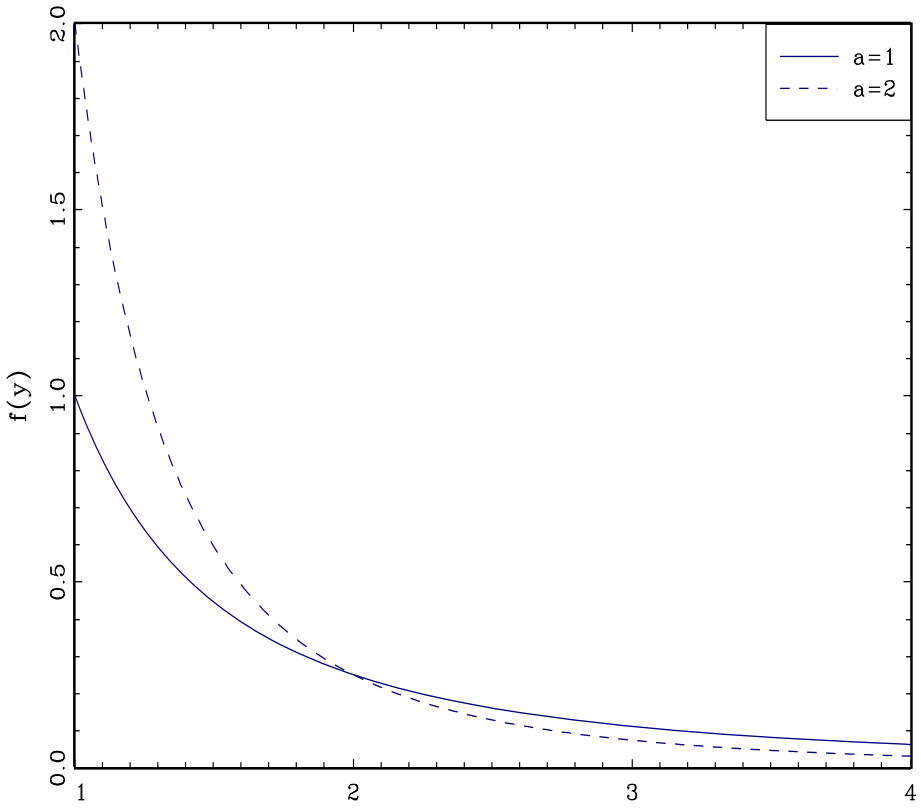


Figure 2.11: Pareto Densities, $a = 1$ and $a = 2$

This connection between tail decay and finite moments is not limited to the Pareto distribution. We can make a similar analysis using a tail bound. Suppose that y has density $f(y)$ which satisfies the bound $f(y) \leq A\left|y\right|^{-a-1}$ for some $A < \infty$ and $a > 0$. Since $f(y)$ is bounded below a scale of a Pareto density, its tail behavior is similarly bounded. This means that for $r < a$

$$\mathbb{E}\left|y\right|^r = \int_{-\infty}^{\infty} \left|y\right|^r f(y)dy \leq \int_{-1}^{1} f(y)dy + 2A\int_{1}^{\infty} y^{r-a-1}dy \leq 1 + \frac{2A}{a-r} < \infty.$$

Thus if the tail of the density declines at the rate $\left|y\right|^{-a-1}$ or faster, then y has finite moments up to (but not including) a. Broadly speaking, the restriction that y has a finite r^{th} moment means that the tail of y's density declines to zero faster than y^{-r-1}. The faster decline of the tail means that the probability of observing an extreme value of y is a more rare event.

We complete this section by adding an alternative representation of expectation in terms of the distribution function.

Theorem 2.30.1 *For any non-negative random variable y*

$$\mathbb{E}\left(y\right) = \int_{0}^{\infty} \Pr\left(y > u\right) du$$

Proof of Theorem 2.30.1: Let $F^*(x) = \Pr\left(y > x\right) = 1 - F(x)$, where $F(x)$ is the distribution function. By integration by parts

$$\mathbb{E}\left(y\right) = \int_{0}^{\infty} ydF(y) = -\int_{0}^{\infty} ydF^*(y) = -\left[yF^*(y)\right]_0^{\infty} + \int_{0}^{\infty} F^*(y)dy = \int_{0}^{\infty} \Pr\left(y > u\right) du$$

as stated. ■

2.31 Moment Generating and Characteristic Functions*

For a random variable z with distribution F its **moment generating function** (MGF) is

$$M(t) = \mathbb{E}\left(\exp\left(tz\right)\right) = \int \exp(tz) dF(z). \tag{2.60}$$

This is also known as the Laplace transformation of the density of z. The MGF is a function of the argument t, and is an alternative representation of the distribution F. It is called the moment generating function since the r^{th} derivative evaluated at zero is the r^{th} uncentered moment. Indeed,

$$M^{(r)}(t) = \mathbb{E}\left(\frac{d^r}{dt^r}\exp(tz)\right) = \mathbb{E}\left(z^r \exp\left(tz\right)\right)$$

and thus the r^{th} derivative at $t = 0$ is

$$M^{(r)}(0) = \mathbb{E}\left(z^r\right).$$

A major limitation with the MGF is that it does not exist for many random variables. Essentially, existence of the integral (2.60) requires the tail of the density of z to decline exponentially. This excludes thick-tailed distributions such as the Pareto.

This limitation is removed if we consider the **characteristic function** (CF) of z, which is defined as

$$C(t) = \mathbb{E}\left(\exp\left(\mathrm{i}tz\right)\right) = \int \exp(\mathrm{i}tz) dF(z)$$

where $\mathrm{i} = \sqrt{-1}$. Like the MGF, the CF is a function of its argument t and is a representation of the distribution function F. The CF is also known as the Fourier transformation of the density of z. Unlike the MGF, the CF exists for all random variables and all values of t since $\exp\left(\mathrm{i}tz\right) = \cos\left(tz\right) + \mathrm{i}\sin\left(tz\right)$ is bounded.

Similarly to the MGF, the r^{th} derivative of the characteristic function evaluated at zero takes the simple form

$$C^{(r)}(0) = \mathrm{i}^r \mathbb{E}\left(z^r\right) \tag{2.61}$$

when such expectations exist. A further connection is that the r^{th} moment is finite if and only if $C^{(r)}(t)$ is continuous at zero.

For random vectors $\boldsymbol{z}$ with distribution F we define the multivariate MGF as

$$M(\boldsymbol{t}) = \mathbb{E}\left(\exp\left(\boldsymbol{t}'\boldsymbol{z}\right)\right) = \int \exp(\boldsymbol{t}'\boldsymbol{z}) dF(\boldsymbol{z}) \tag{2.62}$$

when it exists. Similarly, we define the multivariate CF as

$$C(\boldsymbol{t}) = \mathbb{E}\left(\exp\left(\mathrm{i}\boldsymbol{t}'\boldsymbol{z}\right)\right) = \int \exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{z}) dF(\boldsymbol{z}).$$

2.32 Existence and Uniqueness of the Conditional Expectation*

In Sections 2.3 and 2.6 we defined the conditional mean when the conditioning variables $\boldsymbol{x}$ are discrete and when the variables $(y, \boldsymbol{x})$ have a joint density. We have explored these cases because these are the situations where the conditional mean is easiest to describe and understand. However, the conditional mean exists quite generally without appealing to the properties of either discrete or continuous random variables.

To justify this claim we now present a deep result from probability theory. What it says is that the conditional mean exists for all joint distributions $(y, \boldsymbol{x})$ for which y has a finite mean.

Theorem 2.32.1 ***Existence of the Conditional Mean***
If $\mathbb{E}|y| < \infty$ *then there exists a function* $m(\boldsymbol{x})$ *such that for all sets* $\mathcal{X}$ *for which* $\Pr(\boldsymbol{x} \in \mathcal{X})$ *is defined,*

$$\mathbb{E}(1(\boldsymbol{x} \in \mathcal{X})\, y) = \mathbb{E}(1(\boldsymbol{x} \in \mathcal{X})\, m(\boldsymbol{x})). \tag{2.63}$$

The function $m(\boldsymbol{x})$ *is almost everywhere unique, in the sense that if* $h(\boldsymbol{x})$ *satisfies (2.63), then there is a set* S *such that* $\Pr(S) = 1$ *and* $m(\boldsymbol{x}) = h(\boldsymbol{x})$ *for* $\boldsymbol{x} \in S$*. The function* $m(\boldsymbol{x})$ *is called the* ***conditional mean*** *and is written* $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$.

See, for example, Ash (1972), Theorem 6.3.3.

The conditional mean $m(\boldsymbol{x})$ defined by (2.63) specializes to (2.7) when $(y, \boldsymbol{x})$ have a joint density. The usefulness of definition (2.63) is that Theorem 2.32.1 shows that the conditional mean $m(\boldsymbol{x})$ exists for all finite-mean distributions. This definition allows y to be discrete or continuous, for $\boldsymbol{x}$ to be scalar or vector-valued, and for the components of $\boldsymbol{x}$ to be discrete or continuously distributed.

You may have noticed that Theorem 2.32.1 applies only to sets $\mathcal{X}$ for which $\Pr(\boldsymbol{x} \in \mathcal{X})$ is defined. This is a technical issue –measurability – which we largely side-step in this textbook. Formal probability theory only applies to sets which are measurable – for which probabilities are defined, as it turns out that not all sets satisfy measurability. This is not a practical concern for econometrics, so we defer such distinctions for formal theoretical treatments.

2.33 Identification*

A critical and important issue in structural econometric modeling is identification, meaning that a parameter is uniquely determined by the distribution of the observed variables. It is relatively straightforward in the context of the unconditional and conditional mean, but it is worthwhile to introduce and explore the concept at this point for clarity.

Let F denote the distribution of the observed data, for example the distribution of the pair (y, x). Let $\mathcal{F}$ be a collection of distributions F. Let θ be a parameter of interest (for example, the mean $\mathbb{E}(y)$).

Definition 2.33.1 *A parameter* $\theta \in \mathbb{R}$ *is identified on* $\mathcal{F}$ *if for all* $F \in \mathcal{F}$*, there is a uniquely determined value of* θ.

Equivalently, θ is identified if we can write it as a mapping $\theta = g(F)$ on the set $\mathcal{F}$. The restriction to the set $\mathcal{F}$ is important. Most parameters are identified only on a strict subset of the space of all distributions.

Take, for example, the mean $\mu = \mathbb{E}(y)$. It is uniquely determined if $\mathbb{E}|y| < \infty$, so it is clear that μ is identified for the set $\mathcal{F} = \left\{F : \int_{-\infty}^{\infty} |y|\, dF(y) < \infty\right\}$. However, μ is also well defined when it is either positive or negative infinity. Hence, defining I_1 and I_2 as in (2.57) and (2.58), we can deduce that μ is identified on the set $\mathcal{F} = \{F : \{I_1 < \infty\} \cup \{I_2 < \infty\}\}$.

Next, consider the conditional mean. Theorem 2.32.1 demonstrates that $\mathbb{E}|y| < \infty$ is a sufficient condition for identification.

Theorem 2.33.1 ***Identification of the Conditional Mean***
If $\mathbb{E}|y| < \infty$, the conditional mean $m(\boldsymbol{x}) = \mathbb{E}(y \mid \boldsymbol{x})$ is identified almost everywhere.

It might seem as if identification is a general property for parameters, so long as we exclude degenerate cases. This is true for moments of observed data, but not necessarily for more complicated models. As a case in point, consider the context of censoring. Let y be a random variable with distribution F. Instead of observing y, we observe y^* defined by the censoring rule

$$y^* = \begin{cases} y & \text{if } y \leq \tau \\ \tau & \text{if } y > \tau \end{cases}.$$

That is, y^* is capped at the value τ. A common example is income surveys, where income responses are "top-coded", meaning that incomes above the top code τ are recorded as the top code. The observed variable y^* has distribution

$$F^*(u) = \begin{cases} F(u) & \text{for } u \leq \tau \\ 1 & \text{for } u \geq \tau. \end{cases}$$

We are interested in features of the distribution F not the censored distribution F^*. For example, we are interested in the mean wage $\mu = \mathbb{E}(y)$. The difficulty is that we cannot calculate μ from F^* except in the trivial case where there is no censoring $\Pr(y \geq \tau) = 0$. Thus the mean μ is not generically identified from the censored distribution.

A typical solution to the identification problem is to assume a parametric distribution. For example, let $\mathcal{F}$ be the set of normal distributions $y \sim \mathrm{N}(\mu, \sigma^2)$. It is possible to show that the parameters (μ, σ^2) are identified for all $F \in \mathcal{F}$. That is, if we know that the uncensored distribution is normal, we can uniquely determine the parameters from the censored distribution. This is often called **parametric identification** as identification is restricted to a parametric class of distributions. In modern econometrics this is generally viewed as a second-best solution, as identification has been achieved only through the use of an arbitrary and unverifiable parametric assumption.

A pessimistic conclusion might be that it is impossible to identify parameters of interest from censored data without parametric assumptions. Interestingly, this pessimism is unwarranted. It turns out that we can identify the quantiles q_α of F for $\alpha \leq \Pr(y \leq \tau)$. For example, if 20% of the distribution is censored, we can identify all quantiles for $\alpha \in (0, 0.8)$. This is often called **nonparametric identification** as the parameters are identified without restriction to a parametric class.

What we have learned from this little exercise is that in the context of censored data, moments can only be parametrically identified, while non-censored quantiles are nonparametrically identified. Part of the message is that a study of identification can help focus attention on what can be learned from the data distributions available.

2.34 Technical Proofs*

Proof of Theorem 2.7.1: For convenience, assume that the variables have a joint density $f(y, \boldsymbol{x})$. Since $\mathbb{E}(y \mid \boldsymbol{x})$ is a function of the random vector $\boldsymbol{x}$ only, to calculate its expectation we integrate with respect to the density $f_{\boldsymbol{x}}(\boldsymbol{x})$ of $\boldsymbol{x}$, that is

$$\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x})) = \int_{\mathbb{R}^k} \mathbb{E}(y \mid \boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) \, d\boldsymbol{x}.$$

Substituting in (2.7) and noting that $f_{y|\boldsymbol{x}}(y|\boldsymbol{x}) f_{\boldsymbol{x}}(\boldsymbol{x}) = f(y, \boldsymbol{x})$, we find that the above expression equals

$$\int_{\mathbb{R}^k} \left(\int_{\mathbb{R}} y f_{y|\boldsymbol{x}}(y|\boldsymbol{x})\, dy \right) f_{\boldsymbol{x}}(\boldsymbol{x})\, d\boldsymbol{x} = \int_{\mathbb{R}^k} \int_{\mathbb{R}} y f(y, \boldsymbol{x})\, dy d\boldsymbol{x} = \mathbb{E}(y)$$

the unconditional mean of y. ■

Proof of Theorem 2.7.2: Again assume that the variables have a joint density. It is useful to observe that

$$f(y|\boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{f(y, \boldsymbol{x}_1, \boldsymbol{x}_2)}{f(\boldsymbol{x}_1, \boldsymbol{x}_2)} \frac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f(\boldsymbol{x}_1)} = f(y, \boldsymbol{x}_2|\boldsymbol{x}_1), \tag{2.64}$$

the density of $(y, \boldsymbol{x}_2)$ given $\boldsymbol{x}_1$. Here, we have abused notation and used a single symbol f to denote the various unconditional and conditional densities to reduce notational clutter.

Note that

$$\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{\mathbb{R}} y f(y|\boldsymbol{x}_1, \boldsymbol{x}_2)\, dy. \tag{2.65}$$

Integrating (2.65) with respect to the conditional density of $\boldsymbol{x}_2$ given $\boldsymbol{x}_1$, and applying (2.64) we find that

$$\begin{aligned}
\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) \mid \boldsymbol{x}_1) &= \int_{\mathbb{R}^{k_2}} \mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_2|\boldsymbol{x}_1)\, d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{k_2}} \left(\int_{\mathbb{R}} y f(y|\boldsymbol{x}_1, \boldsymbol{x}_2)\, dy \right) f(\boldsymbol{x}_2|\boldsymbol{x}_1)\, d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f(y|\boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_2|\boldsymbol{x}_1)\, dy d\boldsymbol{x}_2 \\
&= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f(y, \boldsymbol{x}_2|\boldsymbol{x}_1)\, dy d\boldsymbol{x}_2 \\
&= \mathbb{E}(y \mid \boldsymbol{x}_1)
\end{aligned}$$

as stated. ■

Proof of Theorem 2.7.3:

$$\mathbb{E}(g(\boldsymbol{x}) y \mid \boldsymbol{x}) = \int_{\mathbb{R}} g(\boldsymbol{x}) y f_{y|\boldsymbol{x}}(y|\boldsymbol{x})\, dy = g(\boldsymbol{x}) \int_{\mathbb{R}} y f_{y|\boldsymbol{x}}(y|\boldsymbol{x})\, dy = g(\boldsymbol{x}) \mathbb{E}(y \mid \boldsymbol{x})$$

This is (2.8). Equation (2.10) follows by applying the Simple Law of Iterated Expectations to (2.8). ■

Proof of Theorem 2.8.1. Applying Minkowski's Inequality (B.12) to $e = y - m(\boldsymbol{x})$,

$$(\mathbb{E}|e|^r)^{1/r} = (\mathbb{E}|y - m(\boldsymbol{x})|^r)^{1/r} \leq (\mathbb{E}|y|^r)^{1/r} + (\mathbb{E}|m(\boldsymbol{x})|^r)^{1/r} < \infty,$$

where the two parts on the right-hand are finite since $\mathbb{E}|y|^r < \infty$ by assumption and $\mathbb{E}|m(\boldsymbol{x})|^r < \infty$ by the Conditional Expectation Inequality (B.7). The fact that $(\mathbb{E}|e|^r)^{1/r} < \infty$ implies $\mathbb{E}|e|^r < \infty$. ■

Proof of Theorem 2.10.2: The assumption that $\mathbb{E}(y^2) < \infty$ implies that all the conditional expectations below exist.

Using the law of iterated expectations $\mathbb{E}(y \mid \boldsymbol{x}_1) = \mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) \mid \boldsymbol{x}_1)$ and the conditional Jensen's inequality (B.6),

$$(\mathbb{E}(y \mid \boldsymbol{x}_1))^2 = (\mathbb{E}(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) \mid \boldsymbol{x}_1))^2 \leq \mathbb{E}\left((\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2))^2 \mid \boldsymbol{x}_1 \right).$$

Taking unconditional expectations, this implies

$$\mathbb{E}\left(\left(\mathbb{E}(y \mid \boldsymbol{x}_1)\right)^2\right) \leq \mathbb{E}\left(\left(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)\right)^2\right).$$

Similarly,

$$\left(\mathbb{E}(y)\right)^2 \leq \mathbb{E}\left(\left(\mathbb{E}(y \mid \boldsymbol{x}_1)\right)^2\right) \leq \mathbb{E}\left(\left(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)\right)^2\right). \tag{2.66}$$

The variables y, $\mathbb{E}(y \mid \boldsymbol{x}_1)$ and $\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$ all have the same mean $\mathbb{E}(y)$, so the inequality (2.66) implies that the variances are ranked monotonically:

$$0 \leq \operatorname{var}\left(\mathbb{E}(y \mid \boldsymbol{x}_1)\right) \leq \operatorname{var}\left(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)\right). \tag{2.67}$$

Define $e = y - \mathbb{E}(y \mid \boldsymbol{x})$ and $u = \mathbb{E}(y \mid \boldsymbol{x}) - \mu$ so that we have the decomposition

$$y - \mu = e + u.$$

Notice $\mathbb{E}(e \mid \boldsymbol{x}) = 0$ and u is a function of $\boldsymbol{x}$. Thus by the Conditioning Theorem, $\mathbb{E}(eu) = 0$ so e and u are uncorrelated. It follows that

$$\operatorname{var}(y) = \operatorname{var}(e) + \operatorname{var}(u) = \operatorname{var}\left(y - \mathbb{E}(y \mid \boldsymbol{x})\right) + \operatorname{var}\left(\mathbb{E}(y \mid \boldsymbol{x})\right). \tag{2.68}$$

The monotonicity of the variances of the conditional mean (2.67) applied to the variance decomposition (2.68) implies the reverse monotonicity of the variances of the differences, completing the proof. ■

Proof of Theorem 2.18.1. For part 1, by the Expectation Inequality (B.8), (A.24) and Assumption 2.18.1,

$$\left\|\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right\| \leq \mathbb{E}\left\|\boldsymbol{x}\boldsymbol{x}'\right\| = \mathbb{E}\left(\|\boldsymbol{x}\|^2\right) < \infty.$$

Similarly, using the Expectation Inequality (B.8), the Cauchy-Schwarz Inequality (B.10) and Assumption 2.18.1,

$$\|\mathbb{E}(\boldsymbol{x}y)\| \leq \mathbb{E}\|\boldsymbol{x}y\| \leq \left(\mathbb{E}\left(\|\boldsymbol{x}\|^2\right)\right)^{1/2}\left(\mathbb{E}\left(y^2\right)\right)^{1/2} < \infty.$$

Thus the moments $\mathbb{E}(\boldsymbol{x}y)$ and $\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ are finite and well defined.

For part 2, the coefficient $\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y)$ is well defined since $\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}$ exists under Assumption 2.18.1.

Part 3 follows from Definition 2.18.1 and part 2.

For part 4, first note that

$$\begin{aligned}
\mathbb{E}\left(e^2\right) &= \mathbb{E}\left(\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2\right) \\
&= \mathbb{E}\left(y^2\right) - 2\mathbb{E}\left(y\boldsymbol{x}'\right)\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\boldsymbol{\beta} \\
&= \mathbb{E}\left(y^2\right) - 2\mathbb{E}\left(y\boldsymbol{x}'\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y) \\
&\leq \mathbb{E}\left(y^2\right) \\
&< \infty.
\end{aligned}$$

The first inequality holds because $\mathbb{E}\left(y\boldsymbol{x}'\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}(\boldsymbol{x}y)$ is a quadratic form and therefore necessarily non-negative. Second, by the Expectation Inequality (B.8), the Cauchy-Schwarz Inequality (B.10) and Assumption 2.18.1,

$$\|\mathbb{E}(\boldsymbol{x}e)\| \leq \mathbb{E}\|\boldsymbol{x}e\| = \left(\mathbb{E}\left(\|\boldsymbol{x}\|^2\right)\right)^{1/2}\left(\mathbb{E}\left(e^2\right)\right)^{1/2} < \infty.$$

It follows that the expectation $\mathbb{E}(\boldsymbol{x}e)$ is finite, and is zero by the calculation (2.28).

For part 6, Applying Minkowski's Inequality (B.12) to $e = y - \boldsymbol{x}'\boldsymbol{\beta}$,

$$\begin{aligned}
\left(\mathbb{E}\left|e\right|^r\right)^{1/r} &= \left(\mathbb{E}\left|y - \boldsymbol{x}'\boldsymbol{\beta}\right|^r\right)^{1/r} \\
&\leq \left(\mathbb{E}\left|y\right|^r\right)^{1/r} + \left(\mathbb{E}\left|\boldsymbol{x}'\boldsymbol{\beta}\right|^r\right)^{1/r} \\
&\leq \left(\mathbb{E}\left|y\right|^r\right)^{1/r} + \left(\mathbb{E}\left\|\boldsymbol{x}\right\|^r\right)^{1/r} \left\|\boldsymbol{\beta}\right\| \\
&< \infty,
\end{aligned}$$

the final inequality by assumption. ■

Exercises

Exercise 2.1 Find $\mathbb{E}\left(\mathbb{E}\left(\mathbb{E}\left(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3\right) \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) \mid \boldsymbol{x}_1\right)$.

Exercise 2.2 If $\mathbb{E}\left(y \mid x\right) = a + bx$, find $\mathbb{E}\left(yx\right)$ as a function of moments of x.

Exercise 2.3 Prove Theorem 2.8.1.4 using the law of iterated expectations.

Exercise 2.4 Suppose that the random variables y and x only take the values 0 and 1, and have the following joint probability distribution

	$x=0$	$x=1$
$y=0$	.1	.2
$y=1$	.4	.3

Find $\mathbb{E}\left(y \mid x\right)$, $\mathbb{E}\left(y^2 \mid x\right)$ and $\text{var}\left(y \mid x\right)$ for $x = 0$ and $x = 1$.

Exercise 2.5 Show that $\sigma^2(\boldsymbol{x})$ is the best predictor of e^2 given $\boldsymbol{x}$:

(a) Write down the mean-squared error of a predictor $h(\boldsymbol{x})$ for e^2.

(b) What does it mean to be predicting e^2?

(c) Show that $\sigma^2(\boldsymbol{x})$ minimizes the mean-squared error and is thus the best predictor.

Exercise 2.6 Use $y = m(\boldsymbol{x}) + e$ to show that

$$\text{var}\left(y\right) = \text{var}\left(m(\boldsymbol{x})\right) + \sigma^2$$

Exercise 2.7 Show that the conditional variance can be written as

$$\sigma^2(\boldsymbol{x}) = \mathbb{E}\left(y^2 \mid \boldsymbol{x}\right) - \left(\mathbb{E}\left(y \mid \boldsymbol{x}\right)\right)^2.$$

Exercise 2.8 Suppose that y is discrete-valued, taking values only on the non-negative integers, and the conditional distribution of y given $\boldsymbol{x}$ is Poisson:

$$\Pr\left(y = j \mid \boldsymbol{x}\right) = \frac{\exp\left(-\boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\beta}\right)^j}{j!}, \qquad j = 0, 1, 2, ...$$

Compute $\mathbb{E}\left(y \mid \boldsymbol{x}\right)$ and $\text{var}\left(y \mid \boldsymbol{x}\right)$. Does this justify a linear regression model of the form $y = \boldsymbol{x}'\boldsymbol{\beta} + e$?

Hint: If $\Pr\left(y = j\right) = \frac{\exp(-\lambda)\lambda^j}{j!}$, then $\mathbb{E}\left(y\right) = \lambda$ and $\text{var}(y) = \lambda$.

Exercise 2.9 Suppose you have two regressors: x_1 is binary (takes values 0 and 1) and x_2 is categorical with 3 categories (A, B, C). Write $\mathbb{E}\left(y \mid x_1, x_2\right)$ as a linear regression.

Exercise 2.10 True or False. If $y = x\beta + e$, $x \in \mathbb{R}$, and $\mathbb{E}\left(e \mid x\right) = 0$, then $\mathbb{E}\left(x^2 e\right) = 0$.

Exercise 2.11 True or False. If $y = x\beta + e$, $x \in \mathbb{R}$, and $\mathbb{E}\left(xe\right) = 0$, then $\mathbb{E}\left(x^2 e\right) = 0$.

Exercise 2.12 True or False. If $y = \boldsymbol{x}'\boldsymbol{\beta} + e$ and $\mathbb{E}\left(e \mid \boldsymbol{x}\right) = 0$, then e is independent of $\boldsymbol{x}$.

Exercise 2.13 True or False. If $y = \boldsymbol{x}'\boldsymbol{\beta} + e$ and $\mathbb{E}(\boldsymbol{x}e) = \boldsymbol{0}$, then $\mathbb{E}\left(e \mid \boldsymbol{x}\right) = 0$.

Exercise 2.14 True or False. If $y = \boldsymbol{x}'\boldsymbol{\beta} + e$, $\mathbb{E}(e \mid \boldsymbol{x}) = 0$, and $\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \sigma^2$, a constant, then e is independent of $\boldsymbol{x}$.

Exercise 2.15 Consider the intercept-only model $y = \alpha + e$ defined as the best linear predictor. Show that $\alpha = \mathbb{E}(y)$.

Exercise 2.16 Let x and y have the joint density $f(x, y) = \frac{3}{2}\left(x^2 + y^2\right)$ on $0 \leq x \leq 1$, $0 \leq y \leq 1$. Compute the coefficients of the best linear predictor $y = \alpha + \beta x + e$. Compute the conditional mean $m(x) = \mathbb{E}(y \mid x)$. Are the best linear predictor and conditional mean different?

Exercise 2.17 Let x be a random variable with $\mu = \mathbb{E}(x)$ and $\sigma^2 = \text{var}(x)$. Define

$$g\left(x \mid \mu, \sigma^2\right) = \begin{pmatrix} x - \mu \\ (x - \mu)^2 - \sigma^2 \end{pmatrix}.$$

Show that $\mathbb{E}g(x \mid m, s) = 0$ if and only if $m = \mu$ and $s = \sigma^2$.

Exercise 2.18 Suppose that

$$\boldsymbol{x} = \begin{pmatrix} 1 \\ x_2 \\ x_3 \end{pmatrix}$$

and $x_3 = \alpha_1 + \alpha_2 x_2$ is a linear function of x_2.

(a) Show that $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}(\boldsymbol{x}\boldsymbol{x}')$ is not invertible.

(b) Use a linear transformation of $\boldsymbol{x}$ to find an expression for the best linear predictor of y given $\boldsymbol{x}$. (Be explicit, do not just use the generalized inverse formula.)

Exercise 2.19 Show (2.46)-(2.47), namely that for

$$d(\boldsymbol{\beta}) = \mathbb{E}\left(m(\boldsymbol{x}) - \boldsymbol{x}'\boldsymbol{\beta}\right)^2$$

then

$$\begin{aligned} \boldsymbol{\beta} &= \underset{\boldsymbol{b} \in \mathbb{R}^k}{\text{argmin}}\, d(\boldsymbol{b}) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x} m(\boldsymbol{x})\right) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x} y\right). \end{aligned}$$

Hint: To show $\mathbb{E}\left(\boldsymbol{x} m(\boldsymbol{x})\right) = \mathbb{E}\left(\boldsymbol{x} y\right)$ use the law of iterated expectations.

Exercise 2.20 Verify that (2.63) holds with $m(\boldsymbol{x})$ defined in (2.7) when $(y, \boldsymbol{x})$ have a joint density $f(y, \boldsymbol{x})$.

Exercise 2.21 Consider the short and long projections

$$y = x\gamma_1 + e$$

$$y = x\beta_1 + x^2\beta_2 + u$$

(a) Under what condition does $\gamma_1 = \beta_1$?

(b) Now suppose the long projection is

$$y = x\theta_1 + x^3\theta_2 + v$$

Is there a similar condition under which $\gamma_1 = \theta_1$?

Exercise 2.22 Take the homoskedastic model

$$\begin{aligned} y &= \boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{x}_2'\beta_2 + e \\ \mathbb{E}\left(e \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) &= 0 \\ \mathbb{E}\left(e^2 \mid \boldsymbol{x}_1, \boldsymbol{x}_2\right) &= \sigma^2 \\ \mathbb{E}\left(\boldsymbol{x}_2 \mid \boldsymbol{x}_1\right) &= \boldsymbol{\Gamma}\boldsymbol{x}_1 \\ \boldsymbol{\Gamma} &\neq 0 \end{aligned}$$

Suppose the parameter $\boldsymbol{\beta}_1$ is of interest. We know that the exclusion of $\boldsymbol{x}_2$ creates omited variable bias in the projection coefficient on $\boldsymbol{x}_2$. It also changes the equation error. Our question is: what is the effect on the homoskedasticity property of the induced equation error? Does the exclusion of $\boldsymbol{x}_2$ induce heteroskedasticity or not? Be specific.

Chapter 3

The Algebra of Least Squares

3.1 Introduction

In this chapter we introduce the popular least-squares estimator. Most of the discussion will be algebraic, with questions of distribution and inference deferred to later chapters.

3.2 Samples

In Section 2.18 we derived and discussed the best linear predictor of y given $\boldsymbol{x}$ for a pair of random variables $(y, \boldsymbol{x}) \in \mathbb{R} \times \mathbb{R}^k$, and called this the linear projection model. We are now interested in **estimating** the parameters of this model, in particular the projection coefficient

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x}y\right). \tag{3.1}$$

We can estimate $\boldsymbol{\beta}$ from observational data which includes joint measurements on the variables $(y, \boldsymbol{x})$. For example, supposing we are interested in estimating a wage equation, we would use a dataset with observations on wages (or weekly earnings), education, experience (or age), and demographic characteristics (gender, race, location). One possible dataset is the Current Population Survey (CPS), a survey of U.S. households which includes questions on employment, income, education, and demographic characteristics.

Notationally we wish to distinguish observations from the underlying random variables. The convention in econometrics is to denote observations by appending a subscript i which runs from 1 to n, thus the i^{th} observation is $(y_i, \boldsymbol{x}_i)$, and n denotes the sample size. The dataset is then $\{(y_i, \boldsymbol{x}_i);\ i = 1, ..., n\}$. We call this the **sample** or the **observations**.

From the viewpoint of empirical analysis, a dataset is an array of numbers often organized as a table, where the columns of the table correspond to distinct variables and the rows correspond to distinct observations. For empirical analysis, the dataset and observations are fixed in the sense that they are numbers presented to the researcher. For statistical analysis we need to view the dataset as random, or more precisely as a realization of a random process.

In order for the coefficient $\boldsymbol{\beta}$ defined in (3.1) to make sense as defined, the expectations over the random variables $(\boldsymbol{x}, y)$ need to be common across the observations. The most elegant approach to ensure this is to assume that the observations are draws from an identical underlying population F. This is the standard assumption that the observations are identically distributed:

Assumption 3.2.1 *The observations* $\{(y_1, \boldsymbol{x}_1), ..., (y_i, \boldsymbol{x}_i), ..., (y_n, \boldsymbol{x}_n)\}$ *are identically distributed; they are draws from a common distribution* F.

This assumption does not need to be viewed as literally true, rather it is a useful modeling device so that parameters such as $\boldsymbol{\beta}$ are well defined. This assumption should be interpreted as how we view an observation *a priori*, before we actually observe it. If I tell you that we have a sample with $n = 59$ observations set in no particular order, then it makes sense to view two observations, say 17 and 58, as draws from the same distribution. We have no reason to expect anything special about either observation.

In econometric theory, we refer to the underlying common distribution F as the **population**. Some authors prefer the label the **data-generating-process** (DGP). You can think of it as a theoretical concept or an infinitely-large potential population. In contrast we refer to the observations available to us $\{(y_i, \boldsymbol{x}_i);\ i = 1, ..., n\}$ as the **sample** or **dataset**. In some contexts the dataset consists of all potential observations, for example administrative tax records may contain every single taxpayer in a political unit. Even in this case we view the observations as if they are random draws from an underlying infinitely-large population, as this will allow us to apply the tools of statistical theory.

The linear projection model applies to the random observations $(y_i, \boldsymbol{x}_i)$. This means that the probability model for the observations is the same as that described in Section 2.18. We can write the model as

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{3.2}$$

where the linear projection coefficient $\boldsymbol{\beta}$ is defined as

$$\boldsymbol{\beta} = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}}\, S(\boldsymbol{b}), \tag{3.3}$$

the minimizer of the expected squared error

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right), \tag{3.4}$$

and has the explicit solution

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x_i}\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i y_i\right). \tag{3.5}$$

3.3 Moment Estimators

We want to estimate the coefficient $\boldsymbol{\beta}$ defined in (3.5) from the sample of observations. Notice that $\boldsymbol{\beta}$ is written as a function of certain population expectations. In this context an appropriate estimator is the same function of the sample moments. Let's explain this in detail.

To start, suppose that we are interested in the population mean μ of a random variable y_i with distribution function F

$$\mu = \mathbb{E}(y_i) = \int_{-\infty}^{\infty} y dF(y). \tag{3.6}$$

The mean μ is a function of the distribution F as written in (3.6). To estimate μ given a sample $\{y_1, ..., y_n\}$ a natural estimator is the sample mean

$$\widehat{\mu} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Notice that we have written this using two pieces of notation. The notation $\overline{y}$ with the bar on top is conventional for a sample mean. The notation $\widehat{\mu}$ with the hat "^" is conventional in econometrics to denote an estimator of the parameter μ. In this case, the sample mean of y_i is the estimator of μ, so $\widehat{\mu}$ and $\overline{y}$ are the same. The sample mean $\overline{y}$ can be viewed as the natural analog of the population mean (3.6) because $\overline{y}$ equals the expectation (3.6) with respect to the empirical distribution – the discrete distribution which puts weight $1/n$ on each observation y_i. There are many other justifications for $\overline{y}$ as an estimator for μ, we will defer these discussions for now. Suffice it to say

that it is the conventional estimator in the lack of other information about μ or the distribution of y_i.

Now suppose that we are interested in a set of population means of possibly non-linear functions of a random vector $\boldsymbol{y}$, say $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{h}(\boldsymbol{y}_i))$. For example, we may be interested in the first two moments of y_i, $\mathbb{E}(y_i)$ and $\mathbb{E}(y_i^2)$. In this case the natural estimator is the vector of sample means,

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i).$$

For example, $\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2$. This is not really a substantive change. We call $\widehat{\boldsymbol{\mu}}$ the **moment estimator** for $\boldsymbol{\mu}$.

Now suppose that we are interested in a nonlinear function of a set of moments. For example, consider the variance of y

$$\sigma^2 = \operatorname{var}(y_i) = \mathbb{E}(y_i^2) - (\mathbb{E}(y_i))^2.$$

In general, many parameters of interest, say $\boldsymbol{\beta}$, can be written as a function of moments of $\boldsymbol{y}$. Notationally,

$$\begin{aligned}\boldsymbol{\beta} &= \boldsymbol{g}(\boldsymbol{\mu}) \\ \boldsymbol{\mu} &= \mathbb{E}(\boldsymbol{h}(\boldsymbol{y}_i)).\end{aligned}$$

Here, $\boldsymbol{y}_i$ are the random variables, $\boldsymbol{h}(\boldsymbol{y}_i)$ are functions (transformations) of the random variables, and $\boldsymbol{\mu}$ is the mean (expectation) of these functions. $\boldsymbol{\beta}$ is the parameter of interest, and is the (nonlinear) function $\boldsymbol{g}(\cdot)$ of these means.

In this context a natural estimator of $\boldsymbol{\beta}$ is obtained by replacing $\boldsymbol{\mu}$ with $\widehat{\boldsymbol{\mu}}$.

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \boldsymbol{g}(\widehat{\boldsymbol{\mu}}) \\ \widehat{\boldsymbol{\mu}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i)\end{aligned}$$

The estimator $\widehat{\boldsymbol{\beta}}$ is often called a "plug-in" estimator, and sometimes a "substitution" estimator. We typically call $\widehat{\boldsymbol{\beta}}$ a moment, or moment-based, estimator of $\boldsymbol{\beta}$, since it is a natural extension of the moment estimator $\widehat{\boldsymbol{\mu}}$.

Take the example of the variance $\sigma^2 = \operatorname{var}(y_i)$. Its moment estimator is

$$\widehat{\sigma}^2 = \widehat{\mu}_2 - \widehat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2.$$

This is not the only possible estimator for σ^2 (there is the well-known bias-corrected version appropriate for independent observations) but it a straightforward and simple choice.

3.4 Least Squares Estimator

The linear projection coefficient $\boldsymbol{\beta}$ is defined in (3.3) as the minimizer of the expected squared error $S(\boldsymbol{\beta})$ defined in (3.4). For given $\boldsymbol{\beta}$, the expected squared error is the expectation of the squared error $(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2$. The moment estimator of $S(\boldsymbol{\beta})$ is the sample average:

$$\begin{aligned}\widehat{S}(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\ &= \frac{1}{n} SSE(\boldsymbol{\beta})\end{aligned} \tag{3.7}$$

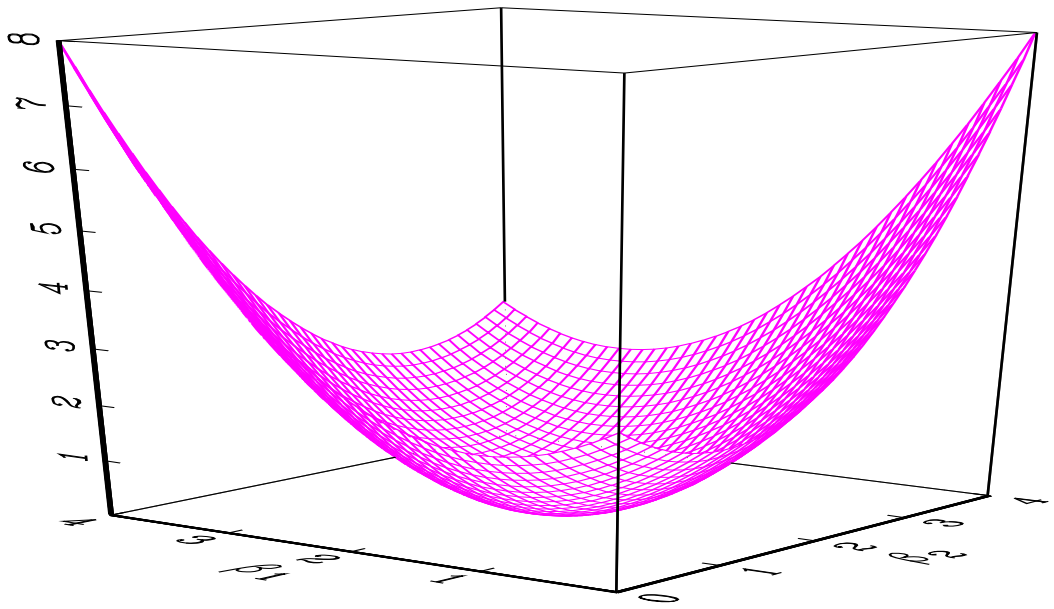


Figure 3.1: Sum-of-Squared Errors Function

where

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \tag{3.8}$$

is called the **sum-of-squared-errors** function.

Since $\widehat{S}(\boldsymbol{\beta})$ is a sample average, we can interpret it as an estimator of the expected squared error $S(\boldsymbol{\beta})$. Examining $\widehat{S}(\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$ is informative about how $S(\boldsymbol{\beta})$ varies with $\boldsymbol{\beta}$. Since the projection coefficient minimizes $S(\boldsymbol{\beta})$, an analog estimator minimizes (3.7):

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^k}{\operatorname{argmin}} \widehat{S}(\boldsymbol{\beta}).$$

Alternatively, as $\widehat{S}(\boldsymbol{\beta})$ is a scale multiple of $SSE(\boldsymbol{\beta})$, we may equivalently define $\widehat{\boldsymbol{\beta}}$ as the minimizer of $SSE_n(\boldsymbol{\beta})$. Hence $\widehat{\boldsymbol{\beta}}$ is commonly called the **least-squares (LS)** estimator of $\boldsymbol{\beta}$. (The estimator is also commonly refered to as the **ordinary least-squares OLS** estimator. For the origin of this label see the historical discussion on Adrien-Marie Legendre below.) Here, as is common in econometrics, we put a hat "^" over the parameter $\boldsymbol{\beta}$ to indicate that $\widehat{\boldsymbol{\beta}}$ is a sample estimate of $\boldsymbol{\beta}$. This is a helpful convention. Just by seeing the symbol $\widehat{\boldsymbol{\beta}}$ we can immediately interpret it as an estimator (because of the hat) of the parameter $\boldsymbol{\beta}$. Sometimes when we want to be explicit about the estimation method, we will write $\widehat{\boldsymbol{\beta}}_{\text{ols}}$ to signify that it is the OLS estimator. It is also common to see the notation $\widehat{\boldsymbol{\beta}}_n$, where the subscript "$n$" indicates that the estimator depends on the sample size n.

It is important to understand the distinction between population parameters such as $\boldsymbol{\beta}$ and sample estimates such as $\widehat{\boldsymbol{\beta}}$. The population parameter $\boldsymbol{\beta}$ is a non-random feature of the population while the sample estimate $\widehat{\boldsymbol{\beta}}$ is a random feature of a random sample. $\boldsymbol{\beta}$ is fixed, while $\widehat{\boldsymbol{\beta}}$ varies across samples.

To visualize the quadratic function $\widehat{S}(\boldsymbol{\beta})$, Figure 3.1 displays an example sum-of-squared errors function $SSE(\boldsymbol{\beta})$ for the case $k = 2$. The least-squares estimator $\widehat{\boldsymbol{\beta}}$ is the the pair $(\widehat{\beta}_1, \widehat{\beta}_2)$ which minimize this function.

3.5 Solving for Least Squares with One Regressor

For simplicity, we start by considering the case $k = 1$ so that the coefficient β is a scalar. Then the sum of squared errors is a simple quadratic

$$\begin{aligned} SSE(\beta) &= \sum_{i=1}^{n} (y_i - x_i\beta)^2 \\ &= \left(\sum_{i=1}^{n} y_i^2\right) - 2\beta\left(\sum_{i=1}^{n} x_i y_i\right) + \beta^2\left(\sum_{i=1}^{n} x_i^2\right). \end{aligned}$$

The OLS estimator $\widehat{\beta}$ minimizes this function. From elementary algebra we know that the minimizer of the quadratic function $a - 2bx + cx^2$ is $x = b/c$. Thus the minimizer of $SSE(\beta)$ is

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}. \tag{3.9}$$

The intercept-only model is the special case $x_i = 1$. In this case we find

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} 1 y_i}{\sum_{i=1}^{n} 1^2} = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}, \tag{3.10}$$

the sample mean of y_i. Here, as is common, we put a bar "$-$" over y to indicate that the quantity is a sample mean. This calculation shows that the OLS estimator in the intercept-only model is the sample mean.

3.6 Solving for Least Squares with Multiple Regressors

We now consider the case with $k \geq 1$ so that the coefficient $\boldsymbol{\beta}$ is a vector.

To solve for $\widehat{\boldsymbol{\beta}}$, expand the SSE function to find

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i^2 - 2\boldsymbol{\beta}' \sum_{i=1}^{n} \boldsymbol{x}_i y_i + \boldsymbol{\beta}' \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \boldsymbol{\beta}.$$

This is a quadratic expression in the vector argument $\boldsymbol{\beta}$. The first-order-condition for minimization of $SSE(\boldsymbol{\beta})$ is

$$0 = \frac{\partial}{\partial \boldsymbol{\beta}} SSE(\widehat{\boldsymbol{\beta}}) = -2\sum_{i=1}^{n} \boldsymbol{x}_i y_i + 2\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}. \tag{3.11}$$

We have written this using a single expression, but it is actually a system of k equations with k unknowns (the elements of $\widehat{\boldsymbol{\beta}}$).

The solution for $\widehat{\boldsymbol{\beta}}$ may be found by solving the system of k equations in (3.11). We can write this solution compactly using matrix algebra. Inverting the $k \times k$ matrix $\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$ we find an explicit formula for the least-squares estimator

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right). \tag{3.12}$$

This is the natural estimator of the best linear projection coefficient $\boldsymbol{\beta}$ defined in (3.3), and can also be called the linear projection estimator.

We see that (3.12) simplifies to the expression (3.9) when $k = 1$. The expression (3.12) is a notationally simple generalization but requires a careful attention to vector and matrix manipulations.

Alternatively, equation (3.5) writes the projection coefficient $\boldsymbol{\beta}$ as an explicit function of the population moments $\boldsymbol{Q}_{xy}$ and $\boldsymbol{Q}_{\boldsymbol{xx}}$. Their moment estimators are the sample moments

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xy}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i$$
$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'.$$

The moment estimator of $\boldsymbol{\beta}$ replaces the population moments in (3.5) with the sample moments:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\widehat{\boldsymbol{Q}}_{\boldsymbol{xy}} \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right)
\end{aligned}$$

which is identical with (3.12).

Least Squares Estimation

Definition 3.6.1 *The* ***least-squares estimator*** $\widehat{\boldsymbol{\beta}}$ *is*

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^k}{\operatorname{argmin}}\, \widehat{S}(\boldsymbol{\beta})$$

where

$$\widehat{S}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

and has the solution

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right).$$

Adrien-Marie Legendre

The method of least-squares was first published in 1805 by the French mathematician Adrien-Marie Legendre (1752-1833). Legendre proposed least-squares as a solution to the algebraic problem of solving a system of equations when the number of equations exceeded the number of unknowns. This was a vexing and common problem in astronomical measurement. As viewed by Legendre, (3.2) is a set of n equations with k unknowns. As the equations cannot be solved exactly, Legendre's goal was to select $\boldsymbol{\beta}$ to make the set of errors as small as possible. He proposed the sum of squared error criterion, and derived the algebraic solution presented above. As he noted, the first-order conditions (3.11) is a system of k equations with k unknowns, which can be solved by "ordinary" methods. Hence the method became known as **Ordinary Least Squares** and to this day we still use the abbreviation OLS to refer to Legendre's estimation method.

3.7 Illustration

We illustrate the least-squares estimator in practice with the data set used to calculate the estimates reported in Chapter 2. This is the March 2009 Current Population Survey, which has extensive information on the U.S. population. This data set is described in more detail in Section 3.19. For this illustration, we use the sub-sample of married (spouse present) black female wage earners with 12 years potential work experience. This sub-sample has 20 observations. Let y_i be log wages and $\boldsymbol{x}_i$ be years of education and an intercept. Then

$$\sum_{i=1}^{n} \boldsymbol{x}_i y_i = \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix},$$

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix},$$

and

$$\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} = \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix}.$$

Thus

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix} \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix}$$

$$= \begin{pmatrix} 0.155 \\ 0.698 \end{pmatrix}. \tag{3.13}$$

We often write the estimated equation using the format

$$\widehat{\log(Wage)} = 0.155\ education + 0.698. \tag{3.14}$$

An interpretation of the estimated equation is that each year of education is associated with a 16% increase in mean wages.

Equation (3.14) is called a **bivariate regression** as there are only two variables. A **multivariate regression** has two or more regressors, and allows a more detailed investigation. Let's take

an example similar to (3.14) but include all levels of experience. This time, we use the sub-sample of single (never married) Asian men, which has 268 observations. Including as regressors years of potential work experience (*experience*) and its square ($experience^2/100$) (we divide by 100 to simplify reporting), we obtain the estimates

$$\widehat{\log(Wage)} = 0.143\ education + 0.036\ experience - 0.071\ experience^2/100 + 0.575. \tag{3.15}$$

These estimates suggest a 14% increase in mean wages per year of education, holding experience constant.

3.8 Least Squares Residuals

As a by-product of estimation, we define the **fitted value**

$$\widehat{y}_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$$

and the **residual**

$$\widehat{e}_i = y_i - \widehat{y}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}. \tag{3.16}$$

Sometimes $\widehat{y}_i$ is called the predicted value, but this is a misleading label. The fitted value $\widehat{y}_i$ is a function of the entire sample, including y_i, and thus cannot be interpreted as a valid prediction of y_i. It is thus more accurate to describe $\widehat{y}_i$ as a *fitted* rather than a *predicted* value.

Note that $y_i = \widehat{y}_i + \widehat{e}_i$ and

$$y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{e}_i. \tag{3.17}$$

We make a distinction between the **error** e_i and the **residual** $\widehat{e}_i$. The error e_i is unobservable while the residual $\widehat{e}_i$ is a by-product of estimation. These two variables are frequently mislabeled, which can cause confusion.

Equation (3.11) implies that

$$\sum_{i=1}^{n} \boldsymbol{x}_i\widehat{e}_i = \mathbf{0}. \tag{3.18}$$

To see this by a direct calculation, using (3.16) and (3.12),

$$\begin{aligned}
\sum_{i=1}^{n} \boldsymbol{x}_i\widehat{e}_i &= \sum_{i=1}^{n} \boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right) \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i y_i \\
&= \mathbf{0}.
\end{aligned}$$

When $\boldsymbol{x}_i$ contains a constant, an implication of (3.18) is

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i = 0. \tag{3.19}$$

Thus the residuals have a sample mean of zero and the sample correlation between the regressors and the residual is zero. These are algebraic results, and hold true for all linear regression estimates.

3.9 Demeaned Regressors

Sometimes it is useful to separate the constant from the other regressors, and write the linear projection equation in the format

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \alpha + e_i$$

where α is the intercept and $\boldsymbol{x}_i$ does not contain a constant. The least-squares estimates and residuals can be written as

$$y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{\alpha} + \widehat{e}_i$$

In this case (3.18) can be written as the equation system

$$\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - \widehat{\alpha}\right) = 0$$

$$\sum_{i=1}^{n}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - \widehat{\alpha}\right) = \mathbf{0}$$

The first equation implies

$$\widehat{\alpha} = \overline{y} - \overline{\boldsymbol{x}}'\widehat{\boldsymbol{\beta}}.$$

Subtracting from the second we obtain

$$\sum_{i=1}^{n}\boldsymbol{x}_i\left((y_i - \overline{y}) - (\boldsymbol{x}_i - \overline{\boldsymbol{x}})'\widehat{\boldsymbol{\beta}}\right) = \mathbf{0}.$$

Solving for $\widehat{\boldsymbol{\beta}}$ we find

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^{n}\boldsymbol{x}_i(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i(y_i - \overline{y})\right) \\ &= \left(\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'\right)^{-1}\left(\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(y_i - \overline{y})\right). \qquad (3.20)\end{aligned}$$

Thus the OLS estimator for the slope coefficients is a regression with demeaned data.

The representation (3.20) is known as the demeaned formula for the least-squares estimator.

3.10 Model in Matrix Notation

For many purposes, including computation, it is convenient to write the model and statistics in matrix notation. The linear equation (2.26) is a system of n equations, one for each observation. We can stack these n equations together as

$$\begin{aligned}y_1 &= \boldsymbol{x}_1'\boldsymbol{\beta} + e_1 \\ y_2 &= \boldsymbol{x}_2'\boldsymbol{\beta} + e_2 \\ &\vdots \\ y_n &= \boldsymbol{x}_n'\boldsymbol{\beta} + e_n.\end{aligned}$$

Now define

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}, \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Observe that $\boldsymbol{y}$ and $\boldsymbol{e}$ are $n \times 1$ vectors, and $\boldsymbol{X}$ is an $n \times k$ matrix. Then the system of n equations can be compactly written in the single equation

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{3.21}$$

Sample sums can be written in matrix notation. For example

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \boldsymbol{X}'\boldsymbol{X}$$
$$\sum_{i=1}^{n} \boldsymbol{x}_i y_i = \boldsymbol{X}'\boldsymbol{y}.$$

Therefore the least-squares estimator can be written as

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}\right). \tag{3.22}$$

The matrix version of (3.17) and estimated version of (3.21) is

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{e}},$$

or equivalently the residual vector is

$$\widehat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Using the residual vector, we can write (3.18) as

$$\boldsymbol{X}'\widehat{\boldsymbol{e}} = \boldsymbol{0}. \tag{3.24}$$

Using matrix notation we have simple expressions for most estimators. This is particularly convenient for computer programming, as most languages allow matrix notation and manipulation.

Important Matrix Expressions

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}\right) \\ \widehat{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ \boldsymbol{X}'\widehat{\boldsymbol{e}} &= \boldsymbol{0}. \end{aligned}$$

Early Use of Matrices

The earliest known treatment of the use of matrix methods to solve simultaneous systems is found in Chapter 8 of the Chinese text *The Nine Chapters on the Mathematical Art*, written by several generations of scholars from the 10th to 2nd century BCE.

3.11 Projection Matrix

Define the matrix

$$\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'.$$

Observe that

$$\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{X}.$$

This is a property of a **projection matrix**. More generally, for any matrix $\boldsymbol{Z}$ which can be written as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Gamma}$ for some matrix $\boldsymbol{\Gamma}$ (we say that $\boldsymbol{Z}$ lies in the **range space** of $\boldsymbol{X}$), then

$$\boldsymbol{P}\boldsymbol{Z} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{Z}.$$

As an important example, if we partition the matrix $\boldsymbol{X}$ into two matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ so that

$$\boldsymbol{X} = \left[\boldsymbol{X}_1 \quad \boldsymbol{X}_2\right],$$

then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$. (See Exercise 3.7.)

The matrix $\boldsymbol{P}$ is **symmetric** ($\boldsymbol{P}' = \boldsymbol{P}$) and **idempotent** ($\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$). (See Section **??**.) To see that it is symmetric,

$$\begin{aligned}
\boldsymbol{P}' &= \left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)' \\
&= \left(\boldsymbol{X}'\right)'\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)'\left(\boldsymbol{X}\right)' \\
&= \boldsymbol{X}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)'\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{X}\left(\left(\boldsymbol{X}\right)'\left(\boldsymbol{X}'\right)'\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{P}.
\end{aligned}$$

To establish that it is idempotent, the fact that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ implies that

$$\begin{aligned}
\boldsymbol{P}\boldsymbol{P} &= \boldsymbol{P}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\
&= \boldsymbol{P}.
\end{aligned}$$

The matrix $\boldsymbol{P}$ has the property that it creates the fitted values in a least-squares regression:

$$\boldsymbol{P}\boldsymbol{y} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{y}}.$$

Because of this property, $\boldsymbol{P}$ is also known as the "hat matrix".

A special example of a projection matrix occurs when $\boldsymbol{X} = \boldsymbol{1}$ is an n-vector of ones. Then

$$\begin{aligned}
\boldsymbol{P}_1 &= \boldsymbol{1}\left(\boldsymbol{1}'\boldsymbol{1}\right)^{-1}\boldsymbol{1}' \\
&= \frac{1}{n}\boldsymbol{1}\boldsymbol{1}'.
\end{aligned}$$

Note that

$$\begin{aligned}
\boldsymbol{P}_1\boldsymbol{y} &= \boldsymbol{1}\left(\boldsymbol{1}'\boldsymbol{1}\right)^{-1}\boldsymbol{1}'\boldsymbol{y} \\
&= \boldsymbol{1}\overline{y}
\end{aligned}$$

creates an n-vector whose elements are the sample mean $\overline{y}$ of y_i.

The i^{th} diagonal element of $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ is

$$h_{ii} = \boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i \tag{3.25}$$

which is called the **leverage** of the i^{th} observation.

Two useful properties of the the matrix $\boldsymbol{P}$ and the leverage values h_{ii} are now summarized.

Theorem 3.11.1

$$\sum_{i=1}^{n} h_{ii} = \operatorname{tr} \boldsymbol{P} = k \tag{3.26}$$

and

$$0 \leq h_{ii} \leq 1. \tag{3.27}$$

To show (3.26),

$$\begin{aligned} \operatorname{tr} \boldsymbol{P} &= \operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right) \\ &= \operatorname{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\right) \\ &= \operatorname{tr}\left(\boldsymbol{I}_k\right) \\ &= k. \end{aligned}$$

See Appendix A.5 for definition and properties of the trace operator. The proof of (3.27) is defered to Section 3.21. One implication is that the rank of $\boldsymbol{P}$ is k.

3.12 Orthogonal Projection

Define

$$\begin{aligned} \boldsymbol{M} &= \boldsymbol{I}_n - \boldsymbol{P} \\ &= \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \end{aligned}$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Note that

$$\boldsymbol{M}\boldsymbol{X} = \left(\boldsymbol{I}_n - \boldsymbol{P}\right)\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{P}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \boldsymbol{0}. \tag{3.28}$$

Thus $\boldsymbol{M}$ and $\boldsymbol{X}$ are orthogonal. We call $\boldsymbol{M}$ an **orthogonal projection matrix**, or more colorfully an **annihilator matrix**, due to the property that for any matrix $\boldsymbol{Z}$ in the range space of $\boldsymbol{X}$ then

$$\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{Z} - \boldsymbol{P}\boldsymbol{Z} = \boldsymbol{0}.$$

For example, $\boldsymbol{M}\boldsymbol{X}_1 = \boldsymbol{0}$ for any subcomponent $\boldsymbol{X}_1$ of $\boldsymbol{X}$, and $\boldsymbol{M}\boldsymbol{P} = \boldsymbol{0}$ (see Exercise 3.7).

The orthogonal projection matrix $\boldsymbol{M}$ has similar properties with $\boldsymbol{P}$, including that $\boldsymbol{M}$ is symmetric ($\boldsymbol{M}' = \boldsymbol{M}$) and idempotent ($\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$). Similarly to (3.26) we can calculate

$$\operatorname{tr} \boldsymbol{M} = n - k. \tag{3.29}$$

(See Exercise 3.9.) One implication is that the rank of $\boldsymbol{M}$ is $n - k$.

While $\boldsymbol{P}$ creates fitted values, $\boldsymbol{M}$ creates least-squares residuals:

$$\boldsymbol{M}\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{e}}. \tag{3.30}$$

As discussed in the previous section, a special example of a projection matrix occurs when $\boldsymbol{X} = \boldsymbol{1}$ is an n-vector of ones, so that $\boldsymbol{P}_1 = \boldsymbol{1}\left(\boldsymbol{1}'\boldsymbol{1}\right)^{-1}\boldsymbol{1}'$. Similarly, set

$$\begin{aligned} \boldsymbol{M}_1 &= \boldsymbol{I}_n - \boldsymbol{P}_1 \\ &= \boldsymbol{I}_n - \boldsymbol{1}\left(\boldsymbol{1}'\boldsymbol{1}\right)^{-1}\boldsymbol{1}'. \end{aligned}$$

While $\boldsymbol{P}_1$ creates a vector of sample means, $\boldsymbol{M}_1$ creates demeaned values:

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{y} - \mathbf{1}\overline{y}.$$

For simplicity we will often write the right-hand-side as $\boldsymbol{y} - \overline{y}$. The i^{th} element is $y_i - \overline{y}$, the **demeaned** value of y_i.

We can also use (3.30) to write an alternative expression for the residual vector. Substituting $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ into $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{y}$ and using $\boldsymbol{M}\boldsymbol{X} = \mathbf{0}$ we find

$$\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}\right) = \boldsymbol{M}\boldsymbol{e} \tag{3.31}$$

which is free of dependence on the regression coefficient $\boldsymbol{\beta}$.

3.13 Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}\left(e_i^2\right)$ is a moment, so a natural estimator is a moment estimator. If e_i were observed we would estimate σ^2 by

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{3.32}$$

However, this is infeasible as e_i is not observed. In this case it is common to take a two-step approach to estimation. The residuals $\widehat{e}_i$ are calculated in the first step, and then we substitute $\widehat{e}_i$ for e_i in expression (3.32) to obtain the feasible estimator

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i^2. \tag{3.33}$$

In matrix notation, we can write (3.32) and (3.33) as

$$\widetilde{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e}$$

and

$$\widehat{\sigma}^2 = n^{-1}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}. \tag{3.34}$$

Recall the expressions $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}\boldsymbol{e}$ from (3.30) and (3.31). Applied to (3.34) we find

$$\begin{aligned}
\widehat{\sigma}^2 &= n^{-1}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}} \\
&= n^{-1}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{M}\boldsymbol{y} \\
&= n^{-1}\boldsymbol{y}'\boldsymbol{M}\boldsymbol{y} \\
&= n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}
\end{aligned} \tag{3.35}$$

the third equality since $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

An interesting implication is that

$$\begin{aligned}
\widetilde{\sigma}^2 - \widehat{\sigma}^2 &= n^{-1}\boldsymbol{e}'\boldsymbol{e} - n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} \\
&= n^{-1}\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} \\
&\geq 0.
\end{aligned}$$

The final inequality holds because $\boldsymbol{P}$ is positive semi-definite and $\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e}$ is a quadratic form. This shows that the feasible estimator $\widehat{\sigma}^2$ is numerically smaller than the idealized estimator (3.32).

3.14 Analysis of Variance

Another way of writing (3.30) is

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} = \widehat{\boldsymbol{y}} + \widehat{\boldsymbol{e}}. \tag{3.36}$$

This decomposition is **orthogonal**, that is

$$\widehat{\boldsymbol{y}}'\widehat{\boldsymbol{e}} = (\boldsymbol{P}\boldsymbol{y})'(\boldsymbol{M}\boldsymbol{y}) = \boldsymbol{y}'\boldsymbol{P}\boldsymbol{M}\boldsymbol{y} = 0.$$

It follows that

$$\boldsymbol{y}'\boldsymbol{y} = \widehat{\boldsymbol{y}}'\widehat{\boldsymbol{y}} + 2\widehat{\boldsymbol{y}}'\widehat{\boldsymbol{e}} + \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}} = \widehat{\boldsymbol{y}}'\widehat{\boldsymbol{y}} + \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$$

or

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \widehat{y}_i^2 + \sum_{i=1}^{n} \widehat{e}_i^2.$$

Subtracting $\bar{y}$ from both sizes of (3.36) we obtain

$$\boldsymbol{y} - \mathbf{1}\overline{y} = \widehat{\boldsymbol{y}} - \mathbf{1}\overline{y} + \widehat{\boldsymbol{e}}.$$

This decomposition is also orthogonal when $\boldsymbol{X}$ contains a constant, as

$$(\widehat{\boldsymbol{y}} - \mathbf{1}\overline{y})'\widehat{\boldsymbol{e}} = \widehat{\boldsymbol{y}}'\widehat{\boldsymbol{e}} - \overline{y}\mathbf{1}'\widehat{\boldsymbol{e}} = 0$$

under (3.19). It follows that

$$(\boldsymbol{y} - \mathbf{1}\overline{y})'(\boldsymbol{y} - \mathbf{1}\overline{y}) = (\hat{\boldsymbol{y}} - \mathbf{1}\overline{y})'(\hat{\boldsymbol{y}} - \mathbf{1}\overline{y}) + \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$$

or

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} \widehat{e}_i^2.$$

This is commonly called the **analysis-of-variance** formula for least squares regression.

A commonly reported statistic is the **coefficient of determination** or **R-squared**:

$$R^2 = \frac{\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n} \widehat{e}_i^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}.$$

It is often described as the fraction of the sample variance of y_i which is explained by the least-squares fit. R^2 is a crude measure of regression fit. We have better measures of fit, but these require a statistical (not just algebraic) analysis and we will return to these issues later. One deficiency with R^2 is that it increases when regressors are added to a regression (see Exercise 3.16) so the "fit" can be always increased by increasing the number of regressors.

3.15 Regression Components

Partition

$$\boldsymbol{X} = [\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}.$$

Then the regression model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}. \tag{3.37}$$

The OLS estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ is obtained by regression of $\boldsymbol{y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and can be written as

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{e}} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{e}}. \tag{3.38}$$

We are interested in algebraic expressions for $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$.

The algebra for the estimator is identical as that for the population coefficients as presented in Section 2.21.

Partition $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}$ as

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2 \\ \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}$$

and similarly $\widehat{\boldsymbol{Q}}_{\boldsymbol{xy}}$

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xy}} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1y} \\ \widehat{\boldsymbol{Q}}_{2y} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{y} \\ \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.4)

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix}^{-1} \overset{def}{=} \begin{bmatrix} \widehat{\boldsymbol{Q}}^{11} & \widehat{\boldsymbol{Q}}^{12} \\ \widehat{\boldsymbol{Q}}^{21} & \widehat{\boldsymbol{Q}}^{22} \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot1}^{-1} \end{bmatrix} \tag{3.39}$$

where $\widehat{\boldsymbol{Q}}_{11\cdot2} = \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}$ and $\widehat{\boldsymbol{Q}}_{22\cdot1} = \widehat{\boldsymbol{Q}}_{22} - \widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{12}$.

Thus

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} \\ &= \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot1}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1y} \\ \widehat{\boldsymbol{Q}}_{2y} \end{bmatrix} \\ &= \begin{pmatrix} \widehat{\boldsymbol{Q}}_{11\cdot2}^{-1}\widehat{\boldsymbol{Q}}_{1y\cdot2} \\ \widehat{\boldsymbol{Q}}_{22\cdot1}^{-1}\widehat{\boldsymbol{Q}}_{2y\cdot1} \end{pmatrix}. \end{aligned}$$

Now

$$\begin{aligned} \widehat{\boldsymbol{Q}}_{11\cdot2} &= \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1 \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1 \end{aligned}$$

where

$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'$$

is the orthogonal projection matrix for $\boldsymbol{X}_2$. Similarly $\widehat{\boldsymbol{Q}}_{22\cdot1} = \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2$ where

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$$

is the orthogonal projection matrix for $\boldsymbol{X}_1$. Also

$$\begin{aligned}\widehat{\boldsymbol{Q}}_{1y\cdot 2} &= \widehat{\boldsymbol{Q}}_{1y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{2y} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{y} - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{y} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}\end{aligned}$$

and $\widehat{\boldsymbol{Q}}_{2y\cdot 1} = \dfrac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}$.

Therefore

$$\widehat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}\right) \tag{3.40}$$

and

$$\widehat{\boldsymbol{\beta}}_2 = \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}\right). \tag{3.41}$$

These are algebraic expressions for the sub-coefficient estimates from (3.38).

3.16 Residual Regression

As first recognized by Frisch and Waugh (1933), expressions (3.40) and (3.41) can be used to show that the least-squares estimators $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ can be found by a two-step regression procedure.

Take (3.41). Since $\boldsymbol{M}_1$ is idempotent, $\boldsymbol{M}_1 = \boldsymbol{M}_1\boldsymbol{M}_1$ and thus

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_2 &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}\right) \\ &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{y}\right) \\ &= \left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{e}}_1\right)\end{aligned}$$

where

$$\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$$

and

$$\widetilde{\boldsymbol{e}}_1 = \boldsymbol{M}_1\boldsymbol{y}.$$

Thus the coefficient estimate $\widehat{\boldsymbol{\beta}}_2$ is algebraically equal to the least-squares regression of $\widetilde{\boldsymbol{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$. Notice that these two are $\boldsymbol{y}$ and $\boldsymbol{X}_2$, respectively, premultiplied by $\boldsymbol{M}_1$. But we know that multiplication by $\boldsymbol{M}_1$ is equivalent to creating least-squares residuals. Therefore $\widetilde{\boldsymbol{e}}_1$ is simply the least-squares residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$, and the columns of $\widetilde{\boldsymbol{X}}_2$ are the least-squares residuals from the regressions of the columns of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$.

We have proven the following theorem.

Theorem 3.16.1 ***Frisch-Waugh-Lovell (FWL)***
In the model (3.37), the OLS estimator of $\boldsymbol{\beta}_2$ and the OLS residuals $\widehat{\boldsymbol{e}}$ may be equivalently computed by either the OLS regression (3.38) or via the following algorithm:

1. *Regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$, obtain residuals $\widetilde{\boldsymbol{e}}_1$;*
2. *Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, obtain residuals $\widetilde{\boldsymbol{X}}_2$;*
3. *Regress $\widetilde{\boldsymbol{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$, obtain OLS estimates $\widehat{\boldsymbol{\beta}}_2$ and residuals $\widehat{\boldsymbol{e}}$.*

In some contexts, the FWL theorem can be used to speed computation, but in most cases there is little computational advantage to using the two-step algorithm.

This result is a direct analogy of the coefficient representation obtained in Section 2.22. The result obtained in that section concerned the population projection coefficients, the result obtained here concern the least-squares estimates. The key message is the same. In the least-squares regression (3.38), the estimated coefficient $\widehat{\boldsymbol{\beta}}_2$ numerically equals the regression of $\boldsymbol{y}$ on the regressors $\boldsymbol{X}_2$, only after the regressors $\boldsymbol{X}_1$ have been linearly projected out. Similarly, the coefficient estimate $\widehat{\boldsymbol{\beta}}_1$ numerically equals the regression of $\boldsymbol{y}$ on the regressors $\boldsymbol{X}_1$, after the regressors $\boldsymbol{X}_2$ have been linearly projected out. This result can be very insightful when interpreting regression coefficients.

A common application of the FWL theorem is the demeaning formula for regression obtained in (3.20).. Partition $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ where $\boldsymbol{X}_1 = \mathbf{1}$ is a vector of ones and $\boldsymbol{X}_2$ is a matrix of observed regressors. In this case,

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'.$$

Observe that

$$\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2 = \boldsymbol{X}_2 - \overline{\boldsymbol{X}}_2$$

and

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{y} - \overline{\boldsymbol{y}}$$

are the "demeaned" variables. The FWL theorem says that $\widehat{\boldsymbol{\beta}}_2$ is the OLS estimate from a regression of $y_i - \overline{y}$ on $\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2$:

$$\widehat{\boldsymbol{\beta}}_2 = \left(\sum_{i=1}^{n}\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)'\right)^{-1}\left(\sum_{i=1}^{n}\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)\left(y_i - \overline{y}\right)\right).$$

This is (3.20).

Ragnar Frisch

Ragnar Frisch (1895-1973) was co-winner with Jan Tinbergen of the first Nobel Memorial Prize in Economic Sciences in 1969 for their work in developing and applying dynamic models for the analysis of economic problems. Frisch made a number of foundational contributions to modern economics beyond the Frisch-Waugh-Lovell Theorem, including formalizing consumer theory, production theory, and business cycle theory.

3.17 Prediction Errors

The least-squares residual $\widehat{e}_i$ are not true prediction errors, as they are constructed based on the full sample including y_i. A proper prediction for y_i should be based on estimates constructed using only the other observations. We can do this by defining the **leave-one-out** OLS estimator of $\boldsymbol{\beta}$ as that obtained from the sample of $n-1$ observations *excluding* the i^{th} observation:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(-i)} &= \left(\frac{1}{n-1}\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j'\right)^{-1}\left(\frac{1}{n-1}\sum_{j\neq i}\boldsymbol{x}_j y_j\right) \\ &= \left(\boldsymbol{X}_{(-i)}'\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}_{(-i)}\boldsymbol{y}_{(-i)}. \end{aligned} \tag{3.42}$$

Here, $\boldsymbol{X}_{(-i)}$ and $\boldsymbol{y}_{(-i)}$ are the data matrices omitting the i^{th} row. The leave-one-out predicted value for y_i is

$$\widetilde{y}_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)},$$

and the **leave-one-out residual** or **prediction error** or **prediction residual** is

$$\widetilde{e}_i = y_i - \widetilde{y}_i.$$

A convenient alternative expression for $\widehat{\boldsymbol{\beta}}_{(-i)}$ (derived in Section 3.21) is

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} - (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\widehat{e}_i \tag{3.43}$$

where h_{ii} are the leverage values as defined in (3.25).

Using (3.43) we can simplify the expression for the prediction error:

$$\begin{aligned}
\widetilde{e}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \\
&= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + (1 - h_{ii})^{-1} \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\widehat{e}_i \\
&= \widehat{e}_i + (1 - h_{ii})^{-1} h_{ii}\widehat{e}_i \\
&= (1 - h_{ii})^{-1} \widehat{e}_i.
\end{aligned} \tag{3.44}$$

To write this in vector notation, define

$$\begin{aligned}
\boldsymbol{M}^* &= (\boldsymbol{I}_n - \text{diag}\{h_{11}, .., h_{nn}\})^{-1} \\
&= \text{diag}\{(1 - h_{11})^{-1}, .., (1 - h_{nn})^{-1}\}.
\end{aligned} \tag{3.45}$$

Then (3.44) is equivalent to

$$\widetilde{\boldsymbol{e}} = \boldsymbol{M}^*\widehat{\boldsymbol{e}}. \tag{3.46}$$

A convenient feature of this expression is that it shows that computation of the full vector of prediction errors $\widetilde{\boldsymbol{e}}$ is based on a simple linear operation, and does not really require n separate estimations.

One use of the prediction errors is to estimate the out-of-sample mean squared error

$$\begin{aligned}
\widetilde{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} \widetilde{e}_i^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} (1 - h_{ii})^{-2} \widehat{e}_i^2.
\end{aligned} \tag{3.47}$$

This is also known as the **sample mean squared prediction error**. Its square root $\widetilde{\sigma} = \sqrt{\widetilde{\sigma}^2}$ is the **prediction standard error**.

3.18 Influential Observations

Another use of the leave-one-out estimator is to investigate the impact of **influential observations**, sometimes called **outliers**. We say that observation i is influential if its omission from the sample induces a substantial change in a parameter estimate of interest.

For illustration, consider Figure 3.2 which shows a scatter plot of random variables (y_i, x_i). The 25 observations shown with the open circles are generated by $x_i \sim U[1, 10]$ and $y_i \sim N(x_i, 4)$. The 26^{th} observation shown with the filled circle is $x_{26} = 9$, $y_{26} = 0$. (Imagine that $y_{26} = 0$ was incorrectly recorded due to a mistaken key entry.) The Figure shows both the least-squares fitted line from the full sample and that obtained after deletion of the 26^{th} observation from the sample. In this example we can see how the 26^{th} observation (the "outlier") greatly tilts the least-squares

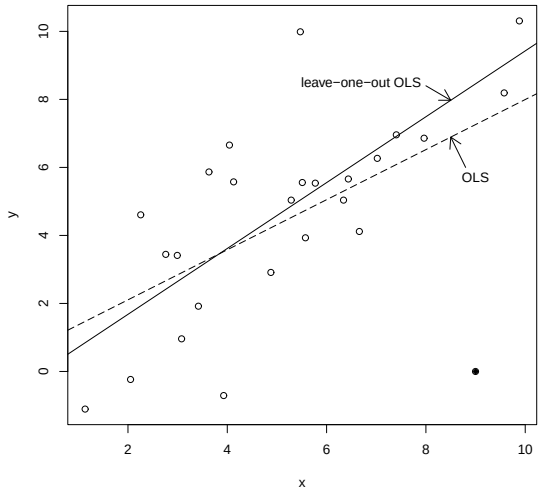


Figure 3.2: Impact of an influential observation on the least-squares estimator

fitted line towards the 26^{th} observation. In fact, the slope coefficient decreases from 0.97 (which is close to the true value of 1.00) to 0.56, which is substantially reduced. Neither y_{26} nor x_{26} are unusual values relative to their marginal distributions, so this outlier would not have been detected from examination of the marginal distributions of the data. The change in the slope coefficient of -0.41 is meaningful and should raise concern to an applied economist.

From (3.43)-(3.44) we know that

$$\begin{aligned}\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)} &= (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \widehat{e}_i \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \widetilde{e}_i. \end{aligned} \tag{3.48}$$

By direct calculation of this quantity for each observation i, we can directly discover if a specific observation i is influential for a coefficient estimate of interest.

For a general assessment, we can focus on the predicted values. The difference between the full-sample and leave-one-out predicted values is

$$\begin{aligned}\widehat{y}_i - \widetilde{y}_i &= \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \\ &= \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \widetilde{e}_i \\ &= h_{ii}\widetilde{e}_i \end{aligned}$$

which is a simple function of the leverage values h_{ii} and prediction errors $\widetilde{e}_i$. Observation i is influential for the predicted value if $|h_{ii}\widetilde{e}_i|$ is large, which requires that both h_{ii} and $|\widetilde{e}_i|$ are large.

One way to think about this is that a large leverage value h_{ii} gives the potential for observation i to be influential. A large h_{ii} means that observation i is unusual in the sense that the regressor $\boldsymbol{x}_i$ is far from its sample mean. We call an observation with large h_{ii} a **leverage point**. A leverage point is not necessarily influential as the latter also requires that the prediction error $\widetilde{e}_i$ is large.

To determine if any individual observations are influential in this sense, several diagnostics have been proposed (some names include DFITS, Cook's Distance, and Welsch Distance). Unfortunately, from a statistical perspective it is difficult to recommend these diagnostics for applications as they are not based on statistical theory. Probably the most relevant measure is the change in the coefficient estimates given in (3.48). The ratio of these changes to the coefficient's standard error is called its DFBETA, and is a postestimation diagnostic available in Stata. While there is no magic threshold, the concern is whether or not an individual observation meaningfully changes an

estimated coefficient of interest. A simple diagnostic for influential observations is to calculate

$$Influence = \max_{1\leq i\leq n} |\widehat{y}_i - \widetilde{y}_i| = \max_{1\leq i\leq n} |h_{ii}\widetilde{e}_i| .$$

This is the largest (absolute) change in the predicted value due to a single observation. If this diagnostic is large relative to the distribution of y_i, it may indicate that that observation is influential.

If an observation is determined to be influential, what should be done? As a common cause of influential observations is data entry error, the influential observations should be examined for evidence that the observation was mis-recorded. Perhaps the observation falls outside of permitted ranges, or some observables are inconsistent (for example, a person is listed as having a job but receives earnings of $0). If it is determined that an observation is incorrectly recorded, then the observation is typically deleted from the sample. This process is often called "cleaning the data". The decisions made in this process involve a fair amount of individual judgment. When this is done it is proper empirical practice to document such choices. (It is useful to keep the source data in its original form, a revised data file after cleaning, and a record describing the revision process. This is especially useful when revising empirical work at a later date.)

It is also possible that an observation is correctly measured, but unusual and influential. In this case it is unclear how to proceed. Some researchers will try to alter the specification to properly model the influential observation. Other researchers will delete the observation from the sample. The motivation for this choice is to prevent the results from being skewed or determined by individual observations, but this practice is viewed skeptically by many researchers who believe it reduces the integrity of reported empirical results.

For an empirical illustration, consider the log wage regression (3.15) for single Asian males. This regression, which has 268 observations, has $Influence = 0.29$. This means that the most influential observation, when deleted, changes the predicted (fitted) value of the dependent variable $\log(Wage)$ by 0.29, or equivalently the wage by 29%. This is a meaningful change and suggests further investigation. We examine the influential observation, and find that its leverage h_{ii} is 0.33, which is disturbingly large. (Recall that the leverage values are all positive and sum to k. One twelfth of the leverage in this sample of 268 observations is contained in just this single observation!) Examining further, we find that this individual is 65 years old with 8 years education, so that his potential experience is 51 years. This is the highest experience in the subsample – the next highest is 41 years. The large leverage is due to his unusual characteristics (very low education and very high experience) within this sample. Essentially, regression (3.15) is attempting to estimate the conditional mean at *experience*= 51 with only one observation, so it is not surprising that this observation determines the fit and is thus influential. A reasonable conclusion is the regression function can only be estimated over a smaller range of *experience*. We restrict the sample to individuals with less than 45 years experience, re-estimate, and obtain the following estimates.

$$\widehat{\log(Wage)} = 0.144\ education + 0.043\ experience - 0.095\ experience^2/100 + 0.531. \tag{3.49}$$

For this regression, we calculate that $Influence = 0.11$, which is greatly reduced relative to the regression (3.15). Comparing (3.49) with (3.15), the slope coefficient for education is essentially unchanged, but the coefficients in experience and its square have slightly increased.

By eliminating the influential observation, equation (3.49) can be viewed as a more robust estimate of the conditional mean for most levels of *experience*. Whether to report (3.15) or (3.49) in an application is largely a matter of judgment.

3.19 CPS Data Set

In this section we describe the data set used in the empirical illustrations.

The Current Population Survey (CPS) is a monthly survey of about 57,000 U.S. households conducted by the Bureau of the Census of the Bureau of Labor Statistics. The CPS is the primary source of information on the labor force characteristics of the U.S. population. The survey covers employment, earnings, educational attainment, income, poverty, health insurance coverage, job experience, voting and registration, computer usage, veteran status, and other variables. Details can be found at www.census.gov/cps and dataferrett.census.gov.

From the March 2009 survey we extracted the individuals with non-allocated variables who were full-time employed (defined as those who had worked at least 36 hours per week for at least 48 weeks the past year), and excluded those in the military. This sample has 50,742 individuals. We extracted 14 variables from the CPS on these individuals and created the data files `cps09mar.dta` (Stata format), `cps09mar.xlsx` (Excel format) and `cps09mar.txt` (text format). The variables are described in the file `cps09mar_description.pdf` All data files are available at http://www.ssc.wisc.edu/~bhansen/econometrics/

3.20 Programming

Most packages allow both interactive programming (where you enter commands one-by-one) and batch programming (where you run a pre-written sequence of commands from a file). Interactive programming can be useful for exploratory analysis, but eventually all work should be executed in batch mode. This is the best way to control and document your work.

Batch programs are text files where each line executes a single command. For Stata, this file needs to have the filename extension ".do", and for MATLAB ".m". For R there is no specific naming requirements, though it is typical to use the extension ".r".

To execute a program file, you type a command within the program.

Stata: `do chapter3` executes the file *chapter3.do*

MATLAB: `run chapter3` executes the file *chapter3.m*

R: `source("chapter3.r")` executes the file *chapter3.r*

When writing batch files, it is useful to include comments for documentation and readability.

We illustrate programming files for Stata, R, and MATLAB, which execute a portion of the empirical illustrations from Sections 3.7 and 3.18.

Stata do File

```
*       Clear memory and load the data
clear
use cps09mar.dta
*       Generate transformations
gen wage=ln(earnings/(hours*week))
gen experience = age - education - 6
gen exp2 = (experience^2)/100
*       Create indicator for subsamples
gen mbf = (race == 2) & (marital <= 2) & (female == 1)
gen sam = (race == 4) & (marital == 7) & (female == 0)
*        Regressions
reg wage education if (mbf == 1) & (experience == 12)
reg wage education experience exp2 if sam == 1
*       Leverage and influence
predict leverage,hat
predict e,residual
gen d=e*leverage/(1-leverage)
summarize d if sam ==1
```

R Program File

```
#       Load the data and create subsamples
dat <- read.table("cps09mar.txt")
experience <- dat[,1]-dat[,4]-6
mbf <- (dat[,11]==2)&(dat[,12]<=2)&(dat[,2]==1)&(experience==12)
sam <- (dat[,11]==4)&(dat[,12]==7)&(dat[,2]==0)
dat1 <- dat[mbf,]
dat2 <- dat[sam,]
#       First regression
y <- as.matrix(log(dat1[,5]/(dat1[,6]*dat1[,7])))
x <- cbind(dat1[,4],matrix(1,nrow(dat1),1))
beta <- solve(t(x)%*%x,t(x)%*%y)
print(beta)
#       Second regression
y <- as.matrix(log(dat2[,5]/(dat2[,6]*dat2[,7])))
experience <- dat2[,1]-dat2[,4]-6
exp2 <- (experience^2)/100
x <- cbind(dat2[,4],experience,exp2,matrix(1,nrow(dat2),1))
beta <- solve(t(x)%*%x,t(x)%*%y)print(beta)
#       Create leverage and influence
e <- y-x%*%beta
leverage <- rowSums(x*(x%*%solve(t(x)%*%x)))
r <- e/(1-leverage)
d <- leverage*e/(1-leverage)
print(max(abs(d)))
```

MATLAB Program File

```
% Load the data and create subsamples
load cps09mar.txt;
dat=cps09mar;
experience=dat(:,1)-dat(:,4)-6;
mbf = (dat(:,11)==2)&(dat(:,12)<=2)&(dat(:,2)==1)&(experience==12);
sam = (dat(:,11)==4)&(dat(:,12)==7)&(dat(:,2)==0);
dat1=dat(mbf,:);
dat2=dat(sam,:);
%       First regression
y=log(dat1(:,5)./(dat1(:,6).*dat1(:,7)));
x=[dat1(:,4),ones(length(dat1),1)];
beta=inv(x'*x)*(x'*y);display(beta);
%       Second regression
y=log(dat2(:,5)./(dat2(:,6).*dat2(:,7)));
experience=dat2(:,1)-dat2(:,4)-6;
exp2 = (experience.^2)/100;
x=[dat2(:,4),experience,exp2,ones(length(dat2),1)];
beta=inv(x'*x)*(x'*y);display(beta);
%       Create leverage and influence
e=y-x*beta;
leverage=sum((x.*(x*inv(x'*x)))')';d=leverage.*e./(1-leverage);
influence=max(abs(d));
display(influence);
```

Instead, to load from an excel file, we can replace the first two lines ('load' and 'dat=') with

```
dat=xlsread('cps09mar.xlsx');
```

3.21 Technical Proofs*

Proof of Theorem 3.11.1, equation (3.27): First, $h_{ii} = \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \geq 0$ since it is a quadratic form and $\boldsymbol{X}'\boldsymbol{X} > 0$. Next, since h_{ii} is the i^{th} diagonal element of the projection matrix $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}$, then

$$h_{ii} = \boldsymbol{s}'\boldsymbol{P}\boldsymbol{s}$$

where

$$\boldsymbol{s} = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$$

is a unit vector with a 1 in the i^{th} place (and zeros elsewhere).

By the spectral decomposition of the idempotent matrix $\boldsymbol{P}$ (see equation (A.10))

$$\boldsymbol{P} = \boldsymbol{B}' \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{B}$$

where $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_n$. Thus letting $\boldsymbol{b} = \boldsymbol{B}\boldsymbol{s}$ denote the i^{th} column of $\boldsymbol{B}$, and partitioning $\boldsymbol{b}' = \begin{pmatrix} \boldsymbol{b}_1' & \boldsymbol{b}_2' \end{pmatrix}$ then

$$\begin{aligned} h_{ii} &= \boldsymbol{s}'\boldsymbol{B}' \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{B}\boldsymbol{s} \\ &= \boldsymbol{b}_1' \begin{bmatrix} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \boldsymbol{b}_1 \\ &= \boldsymbol{b}_1'\boldsymbol{b}_1 \\ &\leq \boldsymbol{b}'\boldsymbol{b} \\ &= 1 \end{aligned}$$

the final equality since $\boldsymbol{b}$ is the i^{th} column of $\boldsymbol{B}$ and $\boldsymbol{B}'\boldsymbol{B} = \boldsymbol{I}_n$. We have shown that $h_{ii} \leq 1$, establishing (3.27). ■

Proof of Equation (3.43). The Sherman–Morrison formula (A.3) from Appendix A.6 states that for nonsingular $\boldsymbol{A}$ and vector $\boldsymbol{b}$

$$\left(\boldsymbol{A} - \boldsymbol{b}\boldsymbol{b}'\right)^{-1} = \boldsymbol{A}^{-1} + \left(1 - \boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1} \boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{A}^{-1}.$$

This implies

$$\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \left(1 - h_{ii}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

and thus

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{(-i)} &= \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i y_i \\ &+ \left(1 - h_{ii}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i\right) \\ &= \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i y_i + \left(1 - h_{ii}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - h_{ii} y_i\right) \\ &= \widehat{\boldsymbol{\beta}} - \left(1 - h_{ii}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \left(\left(1 - h_{ii}\right) y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + h_{ii} y_i\right) \\ &= \widehat{\boldsymbol{\beta}} - \left(1 - h_{ii}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \widehat{e}_i \end{aligned}$$

the third equality making the substitutions $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$ and $h_{ii} = \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i$, and the remainder collecting terms. ■

Exercises

Exercise 3.1 Let y be a random variable with $\mu = \mathbb{E}(y)$ and $\sigma^2 = \text{var}(y)$. Define

$$g\left(y, \mu, \sigma^2\right) = \begin{pmatrix} y - \mu \\ (y-\mu)^2 - \sigma^2 \end{pmatrix}.$$

Let $(\widehat{\mu}, \widehat{\sigma}^2)$ be the values such that $\overline{g}_n(\widehat{\mu}, \widehat{\sigma}^2) = \mathbf{0}$ where $\overline{g}_n(m, s) = n^{-1} \sum_{i=1}^n g\left(y_i, m, s\right)$. Show that $\widehat{\mu}$ and $\widehat{\sigma}^2$ are the sample mean and variance.

Exercise 3.2 Consider the OLS regression of the $n \times 1$ vector $\boldsymbol{y}$ on the $n \times k$ matrix $\boldsymbol{X}$. Consider an alternative set of regressors $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{C}$, where $\boldsymbol{C}$ is a $k \times k$ non-singular matrix. Thus, each column of $\boldsymbol{Z}$ is a mixture of some of the columns of $\boldsymbol{X}$. Compare the OLS estimates and residuals from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$ to the OLS estimates from the regression of $\boldsymbol{y}$ on $\boldsymbol{Z}$.

Exercise 3.3 Using matrix algebra, show $\boldsymbol{X}'\widehat{\boldsymbol{e}} = \mathbf{0}$.

Exercise 3.4 Let $\widehat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$. Find $\boldsymbol{X}_2'\widehat{\boldsymbol{e}}$.

Exercise 3.5 Let $\widehat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}$. Find the OLS coefficient from a regression of $\widehat{\boldsymbol{e}}$ on $\boldsymbol{X}$.

Exercise 3.6 Let $\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Find the OLS coefficient from a regression of $\widehat{\boldsymbol{y}}$ on $\boldsymbol{X}$.

Exercise 3.7 Show that if $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$ and $\boldsymbol{M}\boldsymbol{X}_1 = \mathbf{0}$.

Exercise 3.8 Show that $\boldsymbol{M}$ is idempotent: $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

Exercise 3.9 Show that $\text{tr}\,\boldsymbol{M} = n - k$.

Exercise 3.10 Show that if $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ and $\boldsymbol{X}_1'\boldsymbol{X}_2 = 0$ then $\boldsymbol{P} = \boldsymbol{P}_1 + \boldsymbol{P}_2$.

Exercise 3.11 Show that when $\boldsymbol{X}$ contains a constant, $\dfrac{1}{n}\sum_{i=1}^n \widehat{y}_i = \overline{y}$.

Exercise 3.12 A dummy variable takes on only the values 0 and 1. It is used for categorical data, such as an individual's gender. Let $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ be vectors of 1's and 0's, with the i^{th} element of $\boldsymbol{d}_1$ equaling 1 and that of $\boldsymbol{d}_2$ equaling 0 if the person is a man, and the reverse if the person is a woman. Suppose that there are n_1 men and n_2 women in the sample. Consider fitting the following three equations by OLS

$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.50}$$

$$\boldsymbol{y} = \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.51}$$

$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\phi + \boldsymbol{e} \tag{3.52}$$

Can all three equations (3.50), (3.51), and (3.52) be estimated by OLS? Explain if not.

(a) Compare regressions (3.51) and (3.52). Is one more general than the other? Explain the relationship between the parameters in (3.51) and (3.52).

(b) Compute $\boldsymbol{\iota}'\boldsymbol{d}_1$ and $\boldsymbol{\iota}'\boldsymbol{d}_2$, where $\boldsymbol{\iota}$ is an $n \times 1$ vector of ones.

(c) Letting $\boldsymbol{\alpha} = (\alpha_1 \ \alpha_2)'$, write equation (3.51) as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + e$. Consider the assumption $\mathbb{E}(\boldsymbol{x}_i e_i) = 0$. Is there any content to this assumption in this setting?

Exercise 3.13 Let $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$ be defined as in the previous exercise.

(a) In the OLS regression

$$\boldsymbol{y} = \boldsymbol{d}_1\widehat{\gamma}_1 + \boldsymbol{d}_2\widehat{\gamma}_2 + \widehat{\boldsymbol{u}},$$

show that $\widehat{\gamma}_1$ is the sample mean of the dependent variable among the men of the sample ($\overline{y}_1$), and that $\widehat{\gamma}_2$ is the sample mean among the women ($\overline{y}_2$).

(b) Let $\boldsymbol{X}$ $(n \times k)$ be an additional matrix of regressors. Describe in words the transformations

$$\begin{aligned}\boldsymbol{y}^* &= \boldsymbol{y} - \boldsymbol{d}_1\overline{y}_1 - \boldsymbol{d}_2\overline{y}_2 \\ \boldsymbol{X}^* &= \boldsymbol{X} - \boldsymbol{d}_1\overline{\boldsymbol{x}}_1' - \boldsymbol{d}_2\overline{\boldsymbol{x}}_2'\end{aligned}$$

where $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ are the $k \times 1$ means of the regressors for men and women, respectively.

(c) Compare $\widetilde{\boldsymbol{\beta}}$ from the OLS regression

$$\boldsymbol{y}^* = \boldsymbol{X}^*\widetilde{\boldsymbol{\beta}} + \widetilde{\boldsymbol{e}}$$

with $\widehat{\boldsymbol{\beta}}$ from the OLS regression

$$\boldsymbol{y} = \boldsymbol{d}_1\widehat{\alpha}_1 + \boldsymbol{d}_2\widehat{\alpha}_2 + \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{e}}.$$

Exercise 3.14 Let $\widehat{\boldsymbol{\beta}}_n = \left(\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1}\boldsymbol{X}_n'\boldsymbol{y}_n$ denote the OLS estimate when $\boldsymbol{y}_n$ is $n \times 1$ and $\boldsymbol{X}_n$ is $n \times k$. A new observation $(y_{n+1}, \boldsymbol{x}_{n+1})$ becomes available. Prove that the OLS estimate computed using this additional observation is

$$\widehat{\boldsymbol{\beta}}_{n+1} = \widehat{\boldsymbol{\beta}}_n + \frac{1}{1 + \boldsymbol{x}_{n+1}'\left(\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1}\boldsymbol{x}_{n+1}}\left(\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1}\boldsymbol{x}_{n+1}\left(y_{n+1} - \boldsymbol{x}_{n+1}'\widehat{\boldsymbol{\beta}}_n\right).$$

Exercise 3.15 Prove that R^2 is the square of the sample correlation between $\boldsymbol{y}$ and $\widehat{\boldsymbol{y}}$.

Exercise 3.16 Consider two least-squares regressions

$$\boldsymbol{y} = \boldsymbol{X}_1\widetilde{\boldsymbol{\beta}}_1 + \widetilde{\boldsymbol{e}}$$

and

$$\boldsymbol{y} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{e}}.$$

Let R_1^2 and R_2^2 be the R-squared from the two regressions. Show that $R_2^2 \geq R_1^2$. Is there a case (explain) when there is equality $R_2^2 = R_1^2$?

Exercise 3.17 Show that $\widetilde{\sigma}^2 \geq \widehat{\sigma}^2$. Is equality possible?

Exercise 3.18 For which observations will $\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}}$?

Exercise 3.19 Consider the least-squares regression estimates

$$y_i = x_{1i}\widehat{\beta}_1 + x_{2i}\widehat{\beta}_2 + \widehat{e}_i$$

and the "one regressor at a time" regression estimates

$$y_i = \widetilde{\beta}_1 x_{1i} + \widetilde{e}_{1i} \qquad y_i = \widetilde{\beta}_2 x_{2i} + \widetilde{e}_{2i}$$

Under what condition does $\widetilde{\beta}_1 = \widehat{\beta}_1$ and $\widetilde{\beta}_2 = \widehat{\beta}_2$?

Exercise 3.20 You estimate a least-squares regression

$$y_i = \boldsymbol{x}_{1i}'\widetilde{\boldsymbol{\beta}}_1 + \widetilde{u}_i$$

and then regress the residuals on another set of regressors

$$\widetilde{u}_i = \boldsymbol{x}_{2i}'\widetilde{\boldsymbol{\beta}}_2 + \widetilde{e}_i$$

Does this second regression give you the same estimated coefficients as from estimation of a least-squares regression on both set of regressors?

$$y_i = \boldsymbol{x}_{1i}'\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{x}_{2i}'\widehat{\boldsymbol{\beta}}_2 + \widehat{e}_i$$

In other words, is it true that $\widetilde{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\beta}}_2$? Explain your reasoning.

Exercise 3.21 The data matrix is $(\boldsymbol{y}, \boldsymbol{X})$ with $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$, and consider the transformed regressor matrix $\boldsymbol{Z} = [\boldsymbol{X}_1, \boldsymbol{X}_2 - \boldsymbol{X}_1]$. Suppose you do a least-squares regression of $\boldsymbol{y}$ on $\boldsymbol{X}$, and a least-squares regression of $\boldsymbol{y}$ on $\boldsymbol{Z}$. Let $\widehat{\sigma}^2$ and $\widetilde{\sigma}^2$ denote the residual variance estimates from the two regressions. Give a formula relating $\widehat{\sigma}^2$ and $\widetilde{\sigma}^2$? (Explain your reasoning.)

Exercise 3.22 Use the data set from Section 3.19 and the sub-sample used for equation (3.49) (see Section 3.20) for data construction)

(a) Estimate equation (3.49) and compute the equation R^2 and sum of squared errors.

(b) Re-estimate the slope on education using the residual regression approach. Regress log(Wage) on experience and its square, regress education on experience and its square, and the residuals on the residuals. Report the estimates from this final regression, along with the equation R^2 and sum of squared errors. Does the slope coefficient equal the value in (3.49)? Explain.

(c) Are the R^2 and sum-of-squared errors from parts (a) and (b) equal? Explain.

Exercise 3.23 Estimate equation (3.49) as in part (a) of the previous question. Let $\widehat{e}_i$ be the OLS residual, $\widehat{y}_i$ the predicted value from the regression, x_{1i} be education and x_{2i} be experience. Numerically calculate the following:

(a) $\sum_{i=1}^n \widehat{e}_i$

(b) $\sum_{i=1}^n x_{1i}\widehat{e}_i$

(c) $\sum_{i=1}^n x_{2i}\widehat{e}_i$

(d) $\sum_{i=1}^n x_{1i}^2\widehat{e}_i$

(e) $\sum_{i=1}^n x_{2i}^2\widehat{e}_i$

(f) $\sum_{i=1}^n \widehat{y}_i\widehat{e}_i$

(g) $\sum_{i=1}^n \widehat{e}_i^2$

Are these calculations consistent with the theoretical properties of OLS? Explain.

Exercise 3.24 Use the data set from Section 3.19.

(a) Estimate a log wage regression for the subsample of white male Hispanics. In addition to education, experience, and its square, include a set of binary variables for regions and marital status. For regions, you create dummy variables for Northeast, South and West so that Midwest is the excluded group. For marital status, create variables for married, widowed or divorced, and separated, so that single (never married) is the excluded group.

(b) Repeat this estimation using a different econometric package. Compare your results. Do they agree?

Chapter 4

Least Squares Regression

4.1 Introduction

In this chapter we investigate some finite-sample properties of the least-squares estimator in the linear regression model. In particular, we calculate the finite-sample mean and covariance matrix and propose standard errors for the coefficient estimates.

4.2 Random Sampling

Assumption 3.2.1 specified that the observations have identical distributions. To derive the finite-sample properties of the estimators we will need to additionally specify the dependence structure across the observations.

The simplest context is when the observations are mutually independent, in which case we say that they are **independent and identically distributed**, or **i.i.d.** It is also common to describe iid observations as a **random sample**. Traditionally, random sampling has been the default assumption in cross-section (e.g. survey) contexts. It is quite conveneint as iid sampling leads to straightforward expressions for estimation variance. The assumption seems appropriate (meaning that it should be approximately valid) when samples are small and relatively dispersed. That is, if you randomly sample 1000 people from a large country such as the United States it seems reasonable to model their responses as mutually independent.

Assumption 4.2.1 *The observations* $\{(y_1, \boldsymbol{x}_1), ..., (y_i, \boldsymbol{x}_i), ..., (y_n, \boldsymbol{x}_n)\}$ *are independent and identically distributed.*

For most of this chapter, we will use Assumption 4.2.1 to derive properties of the OLS estimator.

Assumption 4.2.1 means that if you take any two individuals $i \neq j$ in a sample, the values $(y_i, \boldsymbol{x}_i)$ are independent of the values $(y_j, \boldsymbol{x}_j)$ yet have the same distribution. Independence means that the decisions and choices of individual i do not affect the decisions of individual j, and conversely.

This assumption may be violated if individuals in the sample are connected in some way, for example if they are neighbors, members of the same village, classmates at a school, or even firms within a specific industry. In this case, it seems plausible that decisions may be inter-connected and thus mutually dependent rather than independent. Allowing for such interactions complicates inference and requires specialized treatment. A currently popular approach which allows for mutual dependence is known as **clustered dependence**, which assumes that that observations are grouped into "clusters" (for example, schools). We will discuss clustering in more detail in Section 4.20.

4.3 Sample Mean

To start with the simplest setting, we first consider the intercept-only model

$$\begin{aligned} y_i &= \mu + e_i \\ \mathbb{E}(e_i) &= 0. \end{aligned}$$

which is equivalent to the regression model with $k = 1$ and $x_i = 1$. In the intercept model, $\mu = \mathbb{E}(y_i)$ is the mean of y_i. (See Exercise 2.15.) The least-squares estimator $\widehat{\mu} = \overline{y}$ equals the sample mean as shown in equation (3.10).

We now calculate the mean and variance of the estimator $\overline{y}$. Since the sample mean is a linear function of the observations, its expectation is simple to calculate

$$\mathbb{E}(\overline{y}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(y_i) = \mu.$$

This shows that the expected value of the least-squares estimator (the sample mean) equals the projection coefficient (the population mean). An estimator with the property that its expectation equals the parameter it is estimating is called **unbiased**.

Definition 4.3.1 *An estimator $\widehat{\theta}$ for θ is **unbiased** if $\mathbb{E}\left(\widehat{\theta}\right) = \theta$.*

We next calculate the variance of the estimator $\overline{y}$ under Assumption 4.2.1. Making the substitution $y_i = \mu + e_i$ we find

$$\overline{y} - \mu = \frac{1}{n}\sum_{i=1}^{n} e_i.$$

Then

$$\begin{aligned} \operatorname{var}(\overline{y}) &= \mathbb{E}(\overline{y} - \mu)^2 \\ &= \mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} e_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} e_j\right)\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}(e_i e_j) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\ &= \frac{1}{n}\sigma^2. \end{aligned}$$

The second-to-last inequality is because $\mathbb{E}(e_i e_j) = \sigma^2$ for $i = j$ yet $\mathbb{E}(e_i e_j) = 0$ for $i \neq j$ due to independence.

We have shown that $\operatorname{var}(\overline{y}) = \frac{1}{n}\sigma^2$. This is the familiar formula for the variance of the sample mean.

4.4 Linear Regression Model

We now consider the linear regression model. Throughout this chapter we maintain the following.

Assumption 4.4.1 ***Linear Regression Model***
The observations $(y_i, \boldsymbol{x}_i)$ *satisfy the linear regression equation*

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{4.1}$$

$$\mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) = 0. \tag{4.2}$$

The variables have finite second moments

$$\mathbb{E}\left(y_i^2\right) < \infty,$$

$$\mathbb{E}\left\|\boldsymbol{x}_i\right\|^2 < \infty,$$

and an invertible design matrix

$$\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) > 0.$$

We will consider both the general case of heteroskedastic regression, where the conditional variance

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2(\boldsymbol{x}_i) = \sigma_i^2$$

is unrestricted, and the specialized case of homoskedastic regression, where the conditional variance is constant. In the latter case we add the following assumption.

Assumption 4.4.2 ***Homoskedastic Linear Regression Model***
In addition to Assumption 4.4.1,

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2(\boldsymbol{x}_i) = \sigma^2 \tag{4.3}$$

is independent of $\boldsymbol{x}_i$.

4.5 Mean of Least-Squares Estimator

In this section we show that the OLS estimator is unbiased in the linear regression model. This calculation can be done using either summation notation or matrix notation. We will use both.

First take summation notation. Observe that under (4.1)-(4.2)

$$\mathbb{E}\left(y_i \mid \boldsymbol{X}\right) = \mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right) = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{4.4}$$

The first equality states that the conditional expectation of y_i given $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$ only depends on $\boldsymbol{x}_i$, since the observations are independent across i. The second equality is the assumption of a linear conditional mean.

Using definition (3.12), the conditioning theorem, the linearity of expectations, (4.4), and properties of the matrix inverse,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \mathbb{E}\left(\left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^{n} \mathbb{E}\left(\boldsymbol{x}_i y_i \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \mathbb{E}\left(y_i \mid \boldsymbol{X}\right) \\
&= \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \boldsymbol{\beta} \\
&= \boldsymbol{\beta}.
\end{aligned}$$

Now let's show the same result using matrix notation. (4.4) implies

$$\mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \begin{pmatrix} \vdots \\ \mathbb{E}\left(y_i \mid \boldsymbol{X}\right) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \boldsymbol{x}_i' \boldsymbol{\beta} \\ \vdots \end{pmatrix} = \boldsymbol{X}\boldsymbol{\beta}. \tag{4.5}$$

Similarly

$$\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \begin{pmatrix} \vdots \\ \mathbb{E}\left(e_i \mid \boldsymbol{X}\right) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) \\ \vdots \end{pmatrix} = \mathbf{0}. \tag{4.6}$$

Using definition (3.22), the conditioning theorem, the linearity of expectations, (4.5), and the properties of the matrix inverse,

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y} \mid \boldsymbol{X}\right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' \mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}.
\end{aligned}$$

At the risk of belaboring the derivation, another way to calculate the same result is as follows. Insert $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ into the formula (3.22) for $\widehat{\boldsymbol{\beta}}$ to obtain

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}\right)\right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{e}\right) \\
&= \boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{e}.
\end{aligned} \tag{4.7}$$

This is a useful linear decomposition of the estimator $\widehat{\boldsymbol{\beta}}$ into the true parameter $\boldsymbol{\beta}$ and the stochastic component $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{e}$. Once again, we can calculate that

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{e} \mid \boldsymbol{X}\right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' \mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) \\
&= \mathbf{0}.
\end{aligned}$$

Regardless of the method, we have shown that $\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$.

We have shown the following theorem.

Theorem 4.5.1 ***Mean of Least-Squares Estimator***
In the linear regression model (Assumption 4.4.1) and i.i.d. sampling (Assumption 4.2.1)

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta} \tag{4.8}$$

Equation (4.8) says that the estimator $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$, conditional on $\boldsymbol{X}$. This means that the conditional distribution of $\widehat{\boldsymbol{\beta}}$ is centered at $\boldsymbol{\beta}$. By "conditional on $\boldsymbol{X}$" this means that the distribution is unbiased (centered at $\boldsymbol{\beta}$) for any realization of the regressor matrix $\boldsymbol{X}$. In conditional models, we simply refer to this as saying "$\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$".

Strictly speaking, "unbiasedness" is a property of the unconditional distribution. Assuming the unconditional mean is well defined, that is $\mathbb{E}\left\|\widehat{\boldsymbol{\beta}}\right\| < \infty$, then applying the law of iterated expectations, we find that the unconditional mean of $\widehat{\boldsymbol{\beta}}$ is also $\boldsymbol{\beta}$

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbb{E}\left(\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)\right) = \boldsymbol{\beta}. \tag{4.9}$$

4.6 Variance of Least Squares Estimator

In this section we calculate the conditional variance of the OLS estimator.

For any $r \times 1$ random vector $\boldsymbol{Z}$ define the $r \times r$ covariance matrix

$$\begin{aligned} \operatorname{var}(\boldsymbol{Z}) &= \mathbb{E}\left(\left(\boldsymbol{Z} - \mathbb{E}\left(\boldsymbol{Z}\right)\right)\left(\boldsymbol{Z} - \mathbb{E}\left(\boldsymbol{Z}\right)\right)'\right) \\ &= \mathbb{E}\left(\boldsymbol{Z}\boldsymbol{Z}'\right) - \left(\mathbb{E}\left(\boldsymbol{Z}\right)\right)\left(\mathbb{E}\left(\boldsymbol{Z}\right)\right)' \end{aligned}$$

and for any pair $(\boldsymbol{Z}, \boldsymbol{X})$ define the conditional covariance matrix

$$\operatorname{var}(\boldsymbol{Z} \mid \boldsymbol{X}) = \mathbb{E}\left(\left(\boldsymbol{Z} - \mathbb{E}\left(\boldsymbol{Z} \mid \boldsymbol{X}\right)\right)\left(\boldsymbol{Z} - \mathbb{E}\left(\boldsymbol{Z} \mid \boldsymbol{X}\right)\right)' \mid \boldsymbol{X}\right).$$

We define

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} \stackrel{def}{=} \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$$

as the conditional covariance matrix of the regression coefficient estimates. We now derive its form.

The conditional covariance matrix of the $n \times 1$ regression error $\boldsymbol{e}$ is the $n \times n$ matrix

$$\operatorname{var}(\boldsymbol{e} \mid \boldsymbol{X}) = \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right) \stackrel{def}{=} \boldsymbol{D}.$$

The i^{th} diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{X}\right) = \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma_i^2$$

while the ij^{th} off-diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}\left(e_i e_j \mid \boldsymbol{X}\right) = \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right)\mathbb{E}\left(e_j \mid \boldsymbol{x}_j\right) = 0.$$

where the first equality uses independence of the observations (Assumption 1.5.2) and the second is (4.2). Thus $\boldsymbol{D}$ is a diagonal matrix with i^{th} diagonal element σ_i^2:

$$\boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}. \tag{4.10}$$

In the special case of the linear homoskedastic regression model (4.3), then

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma_i^2 = \sigma^2$$

and we have the simplification

$$\boldsymbol{D} = \boldsymbol{I}_n \sigma^2.$$

In general, however, $\boldsymbol{D}$ need not necessarily take this simplified form.

For any $n \times r$ matrix $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{X})$,

$$\operatorname{var}(\boldsymbol{A}'\boldsymbol{y} \mid \boldsymbol{X}) = \operatorname{var}(\boldsymbol{A}'\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A}. \tag{4.11}$$

In particular, we can write $\widehat{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{y}$ where $\boldsymbol{A} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and thus

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \operatorname{var}(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

It is useful to note that

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \sigma_i^2,$$

a weighted version of $\boldsymbol{X}'\boldsymbol{X}$.

In the special case of the linear homoskedastic regression model, $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$, so $\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{X}\sigma^2$, and the variance matrix simplifies to

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2.$$

Theorem 4.6.1 ***Variance of Least-Squares Estimator***
In the linear regression model (Assumption 4.4.1) and i.i.d. sampling (Assumption 4.2.1)

$$\begin{aligned} \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} &= \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \end{aligned} \tag{4.12}$$

where $\boldsymbol{D}$ is defined in (4.10).
In the homoskedastic linear regression model (Assumption 4.4.2) and i.i.d. sampling (Assumption 4.2.1)

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.13}$$

4.7 Gauss-Markov Theorem

Now consider the class of estimators of $\boldsymbol{\beta}$ which are linear functions of the vector $\boldsymbol{y}$, and thus can be written as

$$\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{y}$$

where $\boldsymbol{A}$ is an $n \times k$ function of $\boldsymbol{X}$. As noted before, the least-squares estimator is the special case obtained by setting $\boldsymbol{A} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. What is the best choice of $\boldsymbol{A}$? The Gauss-Markov theorem, which we now present, says that the least-squares estimator is the best choice among linear unbiased estimators when the errors are homoskedastic, in the sense that the least-squares estimator has the smallest variance among all unbiased linear estimators.

To see this, since $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$, then for any linear estimator $\widetilde{\boldsymbol{\beta}} = \boldsymbol{A}'\boldsymbol{y}$ we have

$$\mathbb{E}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{A}'\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{A}'\boldsymbol{X}\boldsymbol{\beta},$$

so $\widetilde{\boldsymbol{\beta}}$ is unbiased if (and only if) $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$. Furthermore, we saw in (4.11) that

$$\operatorname{var}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \operatorname{var}\left(\boldsymbol{A}'\boldsymbol{y} \mid \boldsymbol{X}\right) = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} = \boldsymbol{A}'\boldsymbol{A}\sigma^2$$

the last equality using the homoskedasticity assumption $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$. The "best" unbiased linear estimator is obtained by finding the matrix $\boldsymbol{A}_0$ satisfying $\boldsymbol{A}_0'\boldsymbol{X} = \boldsymbol{I}_k$ such that $\boldsymbol{A}_0'\boldsymbol{A}_0$ is minimized in the positive definite sense, in that for any other matrix $\boldsymbol{A}$ satisfying $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$, then $\boldsymbol{A}'\boldsymbol{A} - \boldsymbol{A}_0'\boldsymbol{A}_0$ is positive semi-definite.

Theorem 4.7.1 ***Gauss-Markov****. In the homoskedastic linear regression model (Assumption 4.4.2) and i.i.d. sampling (Assumption 4.2.1), if $\widetilde{\boldsymbol{\beta}}$ is a linear unbiased estimator of $\boldsymbol{\beta}$ then*

$$\operatorname{var}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) \geq \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

The Gauss-Markov theorem provides a lower bound on the variance matrix of unbiased linear estimators under the assumption of homoskedasticity. It says that no unbiased linear estimator can have a variance matrix smaller (in the positive definite sense) than $\sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Since the variance of the OLS estimator is exactly equal to this bound, this means that the OLS estimator is efficient in the class of linear unbiased estimator. This gives rise to the description of OLS as BLUE, standing for "best linear unbiased estimator". This is is an efficiency justification for the least-squares estimator. The justification is limited because the class of models is restricted to homoskedastic linear regression and the class of potential estimators is restricted to linear unbiased estimators. This latter restriction is particularly unsatisfactory as the theorem leaves open the possibility that a non-linear or biased estimator could have lower mean squared error than the least-squares estimator.

We give a proof of the Gauss-Markov theorem below.

Proof of Theorem 4.7.1.1. Let $\boldsymbol{A}$ be any $n \times k$ function of $\boldsymbol{X}$ such that $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$. The variance of the least-squares estimator is $(\boldsymbol{X}'\boldsymbol{X})^{-1}\sigma^2$ and that of $\boldsymbol{A}'\boldsymbol{y}$ is $\boldsymbol{A}'\boldsymbol{A}\sigma^2$. It is sufficient to show

that the difference $\boldsymbol{A}'\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}$ is positive semi-definite. Set $\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Note that $\boldsymbol{X}'\boldsymbol{C} = \boldsymbol{0}$. Then we calculate that

$$\begin{aligned}\boldsymbol{A}'\boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1} &= \left(\boldsymbol{C} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right)'\left(\boldsymbol{C} + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\right) - (\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \boldsymbol{C}'\boldsymbol{C} + \boldsymbol{C}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C} \\ &\quad + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} - (\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \boldsymbol{C}'\boldsymbol{C}.\end{aligned}$$

The matrix $\boldsymbol{C}'\boldsymbol{C}$ is positive semi-definite (see Appendix A.9) as required.

4.8 Generalized Least Squares

Take the linear regression model in matrix format

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{4.14}$$

Consider a generalized situation where the observation errors are possibly correlated and/or heteroskedastic. Specifically, suppose that

$$\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{0} \tag{4.15}$$

$$\operatorname{var}(\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{\Omega} \tag{4.16}$$

for some $n \times n$ covariance matrix $\boldsymbol{\Omega}$, possibly a function of $\boldsymbol{X}$. This includes the iid sampling framework where $\boldsymbol{\Omega} = \boldsymbol{D}$ but allows for non-diagonal covariance matrices as well.

Under these assumptions, by similar arguments we can calculate the mean and variance of the OLS estimator:

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta} \tag{4.17}$$

$$\operatorname{var}(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{4.18}$$

(see Exercise 4.5).

We have an analog of the Gauss-Markov Theorem.

Theorem 4.8.1 *If (4.15)-(4.16) hold and if $\widetilde{\boldsymbol{\beta}}$ is a linear unbiased estimator of $\boldsymbol{\beta}$ then*

$$\operatorname{var}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) \geq (\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}.$$

We leave the proof for Exercise 4.6.

The theorem provides a lower bound on the variance matrix of unbiased linear estimators. The bound is different from the variance matrix of the OLS estimator except when $\boldsymbol{\Omega} = \boldsymbol{I}_n\sigma^2$. This suggests that we may be able to improve on the OLS estimator.

This is indeed the case when $\boldsymbol{\Omega}$ is known up to scale. That is, suppose that $\boldsymbol{\Omega} = c^2\boldsymbol{\Sigma}$ where $c^2 > 0$ is real and $\boldsymbol{\Sigma}$ is $n \times n$ and known. Take the linear model (4.14) and pre-multiply by $\boldsymbol{\Sigma}^{-1/2}$. This produces the equation

$$\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \widetilde{\boldsymbol{e}}$$

where $\widetilde{\boldsymbol{y}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$, $\widetilde{\boldsymbol{X}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}$, and $\widetilde{\boldsymbol{e}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{e}$. Consider OLS estimation of $\boldsymbol{\beta}$ in this equation

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{\text{gls}} &= \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{y}} \\ &= \left(\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}\right)'\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}\right)\right)^{-1}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{X}\right)'\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{y}. \end{aligned} \tag{4.19}$$

This is called the **Generalized Least Squares** (GLS) estimator of $\boldsymbol{\beta}$.

You can calculate that

$$\mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_{\text{gls}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta} \tag{4.20}$$

$$\operatorname{var}(\widetilde{\boldsymbol{\beta}}_{\text{gls}} \mid \boldsymbol{X}) = \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}. \tag{4.21}$$

This shows that the GLS estimator is unbiased, and has a covariance matrix which equals the lower bound from Theorem 4.8.1. This shows that the lower bound is sharp when $\boldsymbol{\Sigma}$ is known and the GLS is efficient in the class of linear unbiased estimators.

In the linear regression model with independent observations and known conditional variances, where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} = \boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right)$, the GLS estimator takes the form

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{\text{gls}} &= \left(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{y} \\ &= \left(\sum_{i=1}^{n}\sigma_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\sigma_i^{-2}\boldsymbol{x}_i y_i\right).\end{aligned}$$

In practice, the covariance matrix $\boldsymbol{\Omega}$ is unknown, so the GLS estimator as presented here is not feasible. However, the form of the GLS estimator motivates feasible versions, effectively by replacing $\boldsymbol{\Omega}$ with an estimate. We return to this issue in Section 20.2.

4.9 Residuals

What are some properties of the residuals $\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ and prediction errors $\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}$, at least in the context of the linear regression model?

Recall from (3.31) that we can write the residuals in vector notation as

$$\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$$

where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ is the orthogonal projection matrix. Using the properties of conditional expectation

$$\mathbb{E}\left(\widehat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{M}\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{0}$$

and

$$\operatorname{var}\left(\widehat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \operatorname{var}\left(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{M}\operatorname{var}\left(\boldsymbol{e} \mid \boldsymbol{X}\right)\boldsymbol{M} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{M} \tag{4.22}$$

where $\boldsymbol{D}$ is defined in (4.10).

We can simplify this expression under the assumption of conditional homoskedasticity

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2.$$

In this case (4.22) simplifies to

$$\operatorname{var}\left(\widehat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \boldsymbol{M}\sigma^2. \tag{4.23}$$

In particular, for a single observation i, we can find the (conditional) variance of $\widehat{e}_i$ by taking the i^{th} diagonal element of (4.23). Since the i^{th} diagonal element of $\boldsymbol{M}$ is $1 - h_{ii}$ as defined in (3.25) we obtain

$$\operatorname{var}(\widehat{e}_i \mid \boldsymbol{X}) = \mathbb{E}\left(\widehat{e}_i^2 \mid \boldsymbol{X}\right) = (1 - h_{ii})\,\sigma^2. \tag{4.24}$$

As this variance is a function of h_{ii} and hence $\boldsymbol{x}_i$, the residuals $\widehat{e}_i$ are heteroskedastic even if the errors e_i are homoskedastic. Notice as well that this implies $\widehat{e}_i^2$ is a biased estimator of σ^2.

Similarly, recall from (3.46) that the prediction errors $\widetilde{e}_i = (1 - h_{ii})^{-1}\widehat{e}_i$ can be written in vector notation as $\widetilde{\boldsymbol{e}} = \boldsymbol{M}^*\widehat{\boldsymbol{e}}$ where $\boldsymbol{M}^*$ is a diagonal matrix with i^{th} diagonal element $(1 - h_{ii})^{-1}$. Thus $\widetilde{\boldsymbol{e}} = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{e}$. We can calculate that

$$\mathbb{E}(\widetilde{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^*\boldsymbol{M}\mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{0}$$

and

$$\operatorname{var}(\widetilde{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^*\boldsymbol{M}\operatorname{var}(\boldsymbol{e} \mid \boldsymbol{X})\boldsymbol{M}\boldsymbol{M}^* = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^*$$

which simplifies under homoskedasticity to

$$\begin{aligned}\operatorname{var}(\widetilde{\boldsymbol{e}} \mid \boldsymbol{X}) &= \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}\boldsymbol{M}^*\sigma^2 \\ &= \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}^*\sigma^2.\end{aligned}$$

The variance of the i^{th} prediction error is then

$$\begin{aligned}\operatorname{var}(\widetilde{e}_i \mid \boldsymbol{X}) &= \mathbb{E}\left(\widetilde{e}_i^2 \mid \boldsymbol{X}\right) \\ &= (1 - h_{ii})^{-1}(1 - h_{ii})(1 - h_{ii})^{-1}\sigma^2 \\ &= (1 - h_{ii})^{-1}\sigma^2.\end{aligned}$$

A residual with constant conditional variance can be obtained by rescaling. The **standardized residuals** are

$$\bar{e}_i = (1 - h_{ii})^{-1/2}\widehat{e}_i, \tag{4.25}$$

and in vector notation

$$\bar{\boldsymbol{e}} = (\bar{e}_1, ..., \bar{e}_n)' = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{e}. \tag{4.26}$$

From our above calculations, under homoskedasticity,

$$\operatorname{var}(\bar{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{M}^{*1/2}\sigma^2$$

and

$$\operatorname{var}(\bar{e}_i \mid \boldsymbol{X}) = \mathbb{E}\left(\bar{e}_i^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.27}$$

and thus these standardized residuals have the same bias and variance as the original errors when the latter are homoskedastic.

4.10 Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}\left(e_i^2\right)$ can be a parameter of interest even in a heteroskedastic regression or a projection model. σ^2 measures the variation in the "unexplained" part of the regression. Its method of moments estimator (MME) is the sample average of the squared residuals:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2.$$

In the linear regression model we can calculate the mean of $\widehat{\sigma}^2$. From (3.35) and the properties of the trace operator, observe that

$$\widehat{\sigma}^2 = \frac{1}{n}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}'\right).$$

Then

$$\begin{aligned}\mathbb{E}\left(\widehat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\right) \\ &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\right) \\ &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\boldsymbol{D}\right). \end{aligned} \tag{4.28}$$

Adding the assumption of conditional homoskedasticity $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, so that $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$, then (4.28) simplifies to

$$\begin{aligned}\mathbb{E}\left(\widehat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\sigma^2\right) \\ &= \sigma^2\left(\frac{n-k}{n}\right),\end{aligned}$$

the final equality by (3.29). This calculation shows that $\widehat{\sigma}^2$ is biased towards zero. The order of the bias depends on k/n, the ratio of the number of estimated coefficients to the sample size.

Another way to see this is to use (4.24). Note that

$$\begin{aligned}\mathbb{E}\left(\widehat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\widehat{e}_i^2 \mid \boldsymbol{X}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)\sigma^2 \\ &= \left(\frac{n-k}{n}\right)\sigma^2 \end{aligned} \tag{4.29}$$

the last equality using Theorem 3.11.1.

Since the bias takes a scale form, a classic method to obtain an unbiased estimator is by rescaling the estimator. Define

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\widehat{e}_i^2. \tag{4.30}$$

By the above calculation,

$$\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.31}$$

and

$$\mathbb{E}\left(s^2\right) = \sigma^2.$$

Hence the estimator s^2 is unbiased for σ^2. Consequently, s^2 is known as the "bias-corrected estimator" for σ^2 and in empirical practice s^2 is the most widely used estimator for σ^2.

Interestingly, this is not the only method to construct an unbiased estimator for σ^2. An estimator constructed with the standardized residuals $\bar{e}_i$ from (4.25) is

$$\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\bar{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}\widehat{e}_i^2. \tag{4.32}$$

You can show (see Exercise 4.9) that

$$\mathbb{E}\left(\overline{\sigma}^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.33}$$

and thus $\overline{\sigma}^2$ is unbiased for σ^2 (in the homoskedastic linear regression model).

When k/n is small (typically, this occurs when n is large), the estimators $\widehat{\sigma}^2$, s^2 and $\overline{\sigma}^2$ are likely to be close. However, if not then s^2 and $\overline{\sigma}^2$ are generally preferred to $\widehat{\sigma}^2$. Consequently it is best to use one of the bias-corrected variance estimators in applications.

4.11 Mean-Square Forecast Error

A major purpose of estimated regressions is to predict out-of-sample values. Consider an out-of-sample observation $(y_{n+1}, \boldsymbol{x}_{n+1})$ where $\boldsymbol{x}_{n+1}$ is observed but not y_{n+1}. Given the coefficient estimate $\widehat{\boldsymbol{\beta}}$ the standard point estimate of $\mathbb{E}(y_{n+1} \mid \boldsymbol{x}_{n+1}) = \boldsymbol{x}_{n+1}'\boldsymbol{\beta}$ is $\widetilde{y}_{n+1} = \boldsymbol{x}_{n+1}'\widehat{\boldsymbol{\beta}}$. The forecast error is the difference between the actual value y_{n+1} and the point forecast $\widetilde{y}_{n+1}$. This is the forecast error $\widetilde{e}_{n+1} = y_{n+1} - \widetilde{y}_{n+1}$. The mean-squared forecast error (MSFE) is its expected squared value

$$MSFE_n = \mathbb{E}\left(\widetilde{e}_{n+1}^2\right).$$

In the linear regression model, $\widetilde{e}_{n+1} = e_{n+1} - \boldsymbol{x}_{n+1}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$, so

$$\begin{aligned} MSFE_n = \mathbb{E}\left(e_{n+1}^2\right) - 2\mathbb{E}\left(e_{n+1}\boldsymbol{x}_{n+1}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right) \\ + \mathbb{E}\left(\boldsymbol{x}_{n+1}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{x}_{n+1}\right). \end{aligned} \tag{4.34}$$

The first term in (4.34) is σ^2. The second term in (4.34) is zero since $e_{n+1}\boldsymbol{x}_{n+1}'$ is independent of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ and both are mean zero. Using the properties of the trace operator, the third term in (4.34) is

$$\begin{aligned} &\operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right)\right) \\ &= \operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\mathbb{E}\left(\mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mid \boldsymbol{X}\right)\right)\right) \\ &= \operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\mathbb{E}\left(\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\right)\right) \\ &= \mathbb{E}\operatorname{tr}\left(\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\right) \\ &= \mathbb{E}\left(\boldsymbol{x}_{n+1}'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}_{n+1}\right) \end{aligned} \tag{4.35}$$

where we use the fact that $\boldsymbol{x}_{n+1}$ is independent of $\widehat{\boldsymbol{\beta}}$, the definition $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mid \boldsymbol{X}\right)$ and the fact that $\boldsymbol{x}_{n+1}$ is independent of $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$. Thus

$$MSFE_n = \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_{n+1}'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}_{n+1}\right).$$

Under conditional homoskedasticity, this simplifies to

$$MSFE_n = \sigma^2\left(1 + \mathbb{E}\left(\boldsymbol{x}_{n+1}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_{n+1}\right)\right).$$

A simple estimator for the MSFE is obtained by averaging the squared prediction errors (3.47)

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widetilde{e}_i^2$$

where $\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{e}_i(1 - h_{ii})^{-1}$. Indeed, we can calculate that

$$\begin{aligned} \mathbb{E}\left(\widetilde{\sigma}^2\right) &= \mathbb{E}\left(\widetilde{e}_i^2\right) \\ &= \mathbb{E}\left(e_i - \boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)\right)^2 \\ &= \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)'\boldsymbol{x}_i\right). \end{aligned}$$

By a similar calculation as in (4.35) we find

$$\mathbb{E}\left(\widetilde{\sigma}^2\right) = \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_i' \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}_{(-i)}} \boldsymbol{x}_i\right) = MSFE_{n-1}.$$

This is the MSFE based on a sample of size $n-1$, rather than size n. The difference arises because the in-sample prediction errors $\widetilde{e}_i$ for $i \leq n$ are calculated using an effective sample size of $n-1$, while the out-of sample prediction error $\widetilde{e}_{n+1}$ is calculated from a sample with the full n observations. Unless n is very small we should expect $MSFE_{n-1}$ (the MSFE based on $n-1$ observations) to be close to $MSFE_n$ (the MSFE based on n observations). Thus $\widetilde{\sigma}^2$ is a reasonable estimator for $MSFE_n$.

Theorem 4.11.1 ***MSFE***
In the linear regression model (Assumption 4.4.1) and i.i.d. sampling (Assumption 4.2.1)

$$MSFE_n = \mathbb{E}\left(\widetilde{e}_{n+1}^2\right) = \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_{n+1}' \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} \boldsymbol{x}_{n+1}\right)$$

where $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$. *Furthermore,* $\widetilde{\sigma}^2$ *defined in (3.47) is an unbiased estimator of* $MSFE_{n-1}$:

$$\mathbb{E}\left(\widetilde{\sigma}^2\right) = MSFE_{n-1}.$$

4.12 Covariance Matrix Estimation Under Homoskedasticity

For inference, we need an estimate of the covariance matrix $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ of the least-squares estimator. In this section we consider the homoskedastic regression model (Assumption 4.4.2).

Under homoskedasticity, the covariance matrix takes the relatively simple form

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}^0 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \sigma^2$$

which is known up to the unknown scale σ^2. In Section 4.10 we discussed three estimators of σ^2. The most commonly used choice is s^2, leading to the classic covariance matrix estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^0 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} s^2. \tag{4.36}$$

Since s^2 is conditionally unbiased for σ^2, it is simple to calculate that $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^0$ is conditionally unbiased for $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ under the assumption of homoskedasticity:

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^0 \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \mathbb{E}\left(s^2 \mid \boldsymbol{X}\right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \sigma^2 \\
&= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}.
\end{aligned}$$

This was the dominant covariance matrix estimator in applied econometrics for many years, and is still the default method in most regression packages. For example, Stata uses the covariance matrix estimator (4.36) by default in linear regression unless an alternative is specified.

If the estimator (4.36) is used, but the regression error is heteroskedastic, it is possible for $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^0$ to be quite biased for the correct covariance matrix $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. For example,

suppose $k = 1$ and $\sigma_i^2 = x_i^2$ with $\mathbb{E}(x_i) = 0$. The ratio of the true variance of the least-squares estimator to the expectation of the variance estimator is

$$\frac{\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}}{\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0} \mid \boldsymbol{X}\right)} = \frac{\sum_{i=1}^{n} x_i^4}{\sigma^2 \sum_{i=1}^{n} x_i^2} \simeq \frac{\mathbb{E}\left(x_i^4\right)}{\left(\mathbb{E}\left(x_i^2\right)\right)^2} \stackrel{def}{=} \kappa.$$

(Notice that we use the fact that $\sigma_i^2 = x_i^2$ implies $\sigma^2 = \mathbb{E}\left(\sigma_i^2\right) = \mathbb{E}\left(x_i^2\right)$.) The constant κ is the standardized fourth moment (or kurtosis) of the regressor x_i, and can be any number greater than one. For example, if $x_i \sim \mathrm{N}\left(0, \sigma^2\right)$ then $\kappa = 3$, so the true variance $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ is three times larger than the expected homoskedastic estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$. But κ can be much larger. Suppose, for example, that $x_i \sim \chi_1^2 - 1$. In this case $\kappa = 15$, so that the true variance $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ is fifteen times larger than the expected homoskedastic estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$. While this is an extreme and constructed example, the point is that the classic covariance matrix estimator (4.36) may be quite biased when the homoskedasticity assumption fails.

4.13 Covariance Matrix Estimation Under Heteroskedasticity

In the previous section we showed that that the classic covariance matrix estimator can be highly biased if homoskedasticity fails. In this section we show how to construct covariance matrix estimators which do not require homoskedasticity.

Recall that the general form for the covariance matrix is

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

This depends on the unknown matrix $\boldsymbol{D}$ which we can write as

$$\begin{aligned}\boldsymbol{D} &= \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right) \\ &= \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right) \\ &= \mathbb{E}\left(\boldsymbol{D}_0 \mid \boldsymbol{X}\right)\end{aligned}$$

where $\boldsymbol{D}_0 = \operatorname{diag}\left(e_1^2, ..., e_n^2\right)$. Thus $\boldsymbol{D}_0$ is a conditionally unbiased estimator for $\boldsymbol{D}$. If the squared errors e_i^2 were observable, we could construct the unbiased estimator

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{ideal} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}_0\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' e_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned}$$

Indeed,

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{ideal} \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\mathbb{E}\left(e_i^2 \mid \boldsymbol{X}\right)\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\sigma_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\end{aligned}$$

verifying that $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{ideal}$ is unbiased for $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$.

Since the errors e_i^2 are unobserved, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{ideal}$ is not a feasible estimator. However, we can replace the errors e_i with the least-squares residuals $\widehat{e}_i$. Making this substitution we obtain the estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.37}$$

We know, however, that $\widehat{e}_i^2$ is biased towards zero (recall equation (4.24)). To estimate the variance σ^2 the unbiased estimator s^2 scales the moment estimator $\widehat{\sigma}^2$ by $n/(n-k)$. Making the same adjustment we obtain the estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \left(\frac{n}{n-k}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.38}$$

While the scaling by $n/(n-k)$ is *ad hoc*, it is recommended over the unscaled estimator (4.37).

Alternatively, we could use the prediction errors $\widetilde{e}_i$ or the standardized residuals $\bar{e}_i$, yielding the estimators

$$\begin{aligned}\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\widetilde{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\end{aligned} \tag{4.39}$$

and

$$\begin{aligned}\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\bar{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\bar{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned} \tag{4.40}$$

The four estimators $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, and $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ are collectively called **robust**, **heteroskedasticity-consistent**, or **heteroskedasticity-robust** covariance matrix estimators. The estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ was first developed by Eicker (1963) and introduced to econometrics by White (1980), and is sometimes called the **Eicker-White** or **White** covariance matrix estimator. The degree-of-freedom adjustment in $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ was recommended by Hinkley (1977), and is the default robust covariance matrix estimator implemented in Stata. (It is implement by the ",r" option, for example by a regression executed with the command "reg y x, r". In current applied econometric practice, this is the method used by most users.) The estimator $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ was introduced by Horn, Horn and Duncan (1975) (and is implemented using the vce(hc2) option in Stata). The estimator $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ was derived by MacKinnon and White from the jackknife principle, and by Andrews (1991) based on the principle of leave-one-out cross-validation (and is implemented using the vce(hc3) option in Stata).

Since $(1-h_{ii})^{-2} > (1-h_{ii})^{-1} > 1$ it is straightforward to show that

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} < \overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} < \widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \tag{4.41}$$

(See Exercise 4.10). The inequality $\boldsymbol{A} < \boldsymbol{B}$ when applied to matrices means that the matrix $\boldsymbol{B} - \boldsymbol{A}$ is positive definite.

In general, the bias of the covariance matrix estimators is quite complicated, but they greatly simplify under the assumption of homoskedasticity (4.3). For example, using (4.24),

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \mathbb{E}\left(\widetilde{e}_i^2 \mid \boldsymbol{X}\right)\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left(1 - h_{ii}\right)\sigma^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' h_{ii}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\
&< \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\
&= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}.
\end{aligned}$$

This calculation shows that $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ is biased towards zero.

By a similarly calculation (again under homoskedasticity) we can calculate that the estimator $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is unbiased

$$\mathbb{E}\left(\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2. \tag{4.42}$$

(See Exercise 4.11.)

It might seem rather odd to compare the bias of heteroskedasticity-robust estimators under the assumption of homoskedasticity, but it does give us a baseline for comparison.

Another interesting calculation shows that in general (that is, without assuming homoskedasticity) $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is biased away from zero. Indeed, using the definition of the prediction errors (3.44)

$$\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} = e_i - \boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)$$

so

$$\widetilde{e}_i^2 = e_i^2 - 2\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)e_i + \left(\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)\right)^2.$$

Note that e_i and $\widehat{\boldsymbol{\beta}}_{(-i)}$ are functions of non-overlapping observations and are thus independent. Hence $\mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)e_i \mid \boldsymbol{X}\right) = 0$ and

$$\begin{aligned}
\mathbb{E}\left(\widetilde{e}_i^2 \mid \boldsymbol{X}\right) &= \mathbb{E}\left(e_i^2 \mid \boldsymbol{X}\right) - 2\boldsymbol{x}_i'\mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)e_i \mid \boldsymbol{X}\right) + \mathbb{E}\left(\left(\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)\right)^2 \mid \boldsymbol{X}\right) \\
&= \sigma_i^2 + \mathbb{E}\left(\left(\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)\right)^2 \mid \boldsymbol{X}\right) \\
&\geq \sigma_i^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}\left(\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \mathbb{E}\left(\widetilde{e}_i^2 \mid \boldsymbol{X}\right)\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&\geq \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \sigma_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}.
\end{aligned}$$

This means that $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is conservative in the sense that it is weakly larger (in expectation) than the correct variance for any realization of $\boldsymbol{X}$.

We have introduced five covariance matrix estimators, $\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}}$, $\widehat{\boldsymbol{V}}^W_{\widehat{\boldsymbol{\beta}}}$, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, and $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$. Which should you use? The classic estimator $\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}}$ is typically a poor choice, as it is only valid under the unlikely homoskedasticity restriction. For this reason it is not typically used in contemporary econometric research. Unfortunately, standard regression packages set their default choice as $\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}}$, so users must intentionally select a robust covariance matrix estimator.

Of the four robust estimators, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is the most commonly used as it is the default robust covariance matrix option in Stata. However, $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ may be the preferred choice since it is conservative for any $\boldsymbol{X}$. As $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is simple to implement, this should not be a barrier.

Halbert L. White

Hal White (1950-2012) of the United States was an influential econometrician of recent years. His 1980 paper on heteroskedasticity-consistent covariance matrix estimation for many years has been the most cited paper in economics. His research was central to the movement to view econometric models as approximations, and to the drive for increased mathematical rigor in the discipline. In addition to being a highly prolific and influential scholar, he also co-founded the economic consulting firm Bates White.

4.14 Standard Errors

A variance estimator such as $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is an estimate of the variance of the distribution of $\widehat{\boldsymbol{\beta}}$. A more easily interpretable measure of spread is its square root – the standard deviation. This is so important when discussing the distribution of parameter estimates, we have a special name for estimates of their standard deviation.

Definition 4.14.1 *A **standard error** $s(\widehat{\beta})$ for a real-valued estimator $\widehat{\beta}$ is an estimate of the standard deviation of the distribution of $\widehat{\beta}$.*

When $\boldsymbol{\beta}$ is a vector with estimate $\widehat{\boldsymbol{\beta}}$ and covariance matrix estimate $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, standard errors for individual elements are the square roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$. That is,

$$s(\widehat{\beta}_j) = \sqrt{\widehat{\boldsymbol{V}}_{\widehat{\beta}_j}} = \sqrt{\left[\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\right]_{jj}}.$$

When the classical covariance matrix estimate (4.36) is used, the standard error takes the particularly simple form

$$s(\widehat{\beta}_j) = s\sqrt{\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}}. \tag{4.43}$$

As we discussed in the previous section, there are multiple possible covariance matrix estimators, so standard errors are not unique. It is therefore important to understand what formula and method is used by an author when studying their work. It is also important to understand that a particular standard error may be relevant under one set of model assumptions, but not under another set of assumptions.

To illustrate, we return to the log wage regression (3.14) of Section 3.7. We calculate that $s^2 = 0.160$. Therefore the homoskedastic covariance matrix estimate is

$$\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}} = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} 0.160 = \begin{pmatrix} 0.002 & -0.031 \\ -0.031 & 0.499 \end{pmatrix}.$$

We also calculate that

$$\sum_{i=1}^{n} (1 - h_{ii})^{-1} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2 = \begin{pmatrix} 763.26 & 48.513 \\ 48.513 & 3.1078 \end{pmatrix}.$$

Therefore the Horn-Horn-Duncan covariance matrix estimate is

$$\begin{aligned} \overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} &= \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \begin{pmatrix} 763.26 & 48.513 \\ 48.513 & 3.1078 \end{pmatrix} \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.001 & -0.015 \\ -0.015 & 0.243 \end{pmatrix}. \end{aligned} \tag{4.44}$$

The standard errors are the square roots of the diagonal elements of these matrices. A conventional format to write the estimated equation with standard errors is

$$\begin{array}{cccc} \widehat{\log(Wage)} = & 0.155 & Education + & 0.698 \\ & (0.031) & & (0.493) \end{array}.$$

Alternatively, standard errors could be calculated using the other formulae. We report the different standard errors in the following table.

	Education	Intercept
Homoskedastic (4.36)	0.045	0.707
White (4.37)	0.029	0.461
Scaled White (4.38)	0.030	0.486
Andrews (4.39)	0.033	0.527
Horn-Horn-Duncan (4.40)	0.031	0.493

The homoskedastic standard errors are noticeably different (larger, in this case) than the others, but the four robust standard errors are quite close to one another.

4.15 Covariance Matrix Estimation with Sparse Dummy Variables

The heteroskedasticity-robust covariance matrix estimators can be quite imprecise in some contexts. One is in the presence of **sparse dummy variables** – when a dummy variable only takes the value 1 or 0 for very few observations. In these contexts one component of the variance matrix is estimated on just those few observations and thus will be imprecise. This is effectively hidden from the user.

To see the problem, let d_{1i} be a dummy variable (takes on the values 1 and 0) for "group 1" and let $d_{2i} = 1 - d_{1i}$ be the complement for "group 2" Consider the dummy-only regression

$$y_i = \beta_1 d_{1i} + \beta_2 d_{2i} + e_i$$

which excludes the intercept for identification. The number of observations in the two "groups" are $n_1 = \sum_{i-1}^{n} d_{1i}$ and $n_2 = \sum_{i-1}^{n} d_{2i}$. The least-squares estimates for β_1 and β_2 are the averages

within the two groups. We say the design is sparse if either n_1 or n_2 is small. One implication is that the coefficient for the small group will be imprecisely estimated.

An extreme situation is when $n_1 = 1$, thus group 1 has only a single observation. This would be unlikely to occur intentionally, but is actually remarkably likely when a large number of interactions are included in a regression. In this context, the least-squares estimate for β_1 is $\widehat{\beta}_1 = y_1$, where for simplicity we have assumed that the first observation is the one for which $d_{1i} = 1$. This means that the corresponding residual is $\widehat{e}_1 = 0$.

The implication for covariance matrix estimation is rather unpleasant. The White estimator is

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} = \begin{pmatrix} 0 & 0 \\ 0 & \frac{\widehat{\sigma}^2}{n-1} \end{pmatrix}$$

where $\widehat{\sigma}^2$ is a variance estimator computed with all observations excluding the first. The covariance matrix $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ is singular, and in particular produces the standard error $s(\widehat{\beta}_1) = 0$! That is, the standard regression package will print out a standard error of 0 for the least-precisely estimated coefficient!

The reason is that the estimator is effectively estimating the variance of $\widehat{\beta}_1$ from a single observation. The point estimate of a variance from a single observation is 0. Essentially, while it is impossible to estimate a variance from a single observation the standard formula gives a misleadingly precise answer.

In most practical regressions, estimated standard errors will not be zero as we typically estimate models with an omitted dummy category and an intercept. What are the implications? In this case, while the reported "standard errors" are non-zero, the covariance matrix estimator itself is singular. This means that there is a linear combination of the estimates with a zero estimated variance. This is generally troubling as this situation is largely hidden from the user.

This problem does not arise if the homoskedastic form of the covariance matrix estimate is used. In the above example, the estimate is

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0} = \begin{pmatrix} s^2 & 0 \\ 0 & \frac{s^2}{n-1} \end{pmatrix}.$$

Consequently, in models with sparse dummy variable designs, it may be prudent to use (or at least check) the homoskedastic standard error formulae.

In general, users should be cautious about regression results when dummy variables (and interactions of dummy variables) are sparse.

4.16 Computation

We illustrate methods to compute standard errors for equation (3.15) extending the code of Section 3.20.

Stata do File (continued)

```
*       Homoskedastic formula (4.36):
reg wage education experience exp2 if (mnwf == 1)
*       Scaled White formula (4.38):
reg wage education experience exp2 if (mnwf == 1), r
*       Andrews formula (4.39):
reg wage education experience exp2 if (mnwf == 1), vce(hc3)
*       Horn-Horn-Duncan formula (4.40):
reg wage education experience exp2 if (mnwf == 1), vce(hc2)
```

R Program File (continued)

```
n <- nrow(y)
k <- ncol(x)
a <- n/(n-k)
sig2 <- (t(e) %*% e)/(n-k)
u1 <- x*(e%*%matrix(1,1,k))
u2 <- x*((e/(1-leverage))%*%matrix(1,1,k))
u3 <- x*((e/sqrt(1-leverage))%*%matrix(1,1,k))
v0 <- xx*sig2
xx <- solve(t(x)%*%x)
v1 <- xx %*% (t(u1)%*%u1) %*% xx
v1a <- a * xx %*% (t(u1)%*%u1) %*% xx
v2 <- xx %*% (t(u2)%*%u2) %*% xx
v3 <- xx %*% (t(u3)%*%u3) %*% xx
s0 <- sqrt(diag(v0))          # Homoskedastic formula
s1 <- sqrt(diag(v1))          # White formula
s1a <- sqrt(diag(v1a))        # Scaled White formula
s2 <- sqrt(diag(v2))          # Andrews formula
s3 <- sqrt(diag(v3))          # Horn-Horn-Duncan formula
```

MATLAB Program File (continued)

```
[n,k]=size(x);
a=n/(n-k);
sig2=(e'*e)/(n-k);
u1=x.*(e*ones(1,k));
u2=x.*((e./(1-leverage))*ones(1,k));u3=x.*((e./sqrt(1-
leverage))*ones(1,k));
xx=inv(x'*x);
v0=xx*sig2;
v1=xx*(u1'*u1)*xx;
v1a=a*xx*(u1'*u1)*xx;
v2=xx*(u2'*u2)*xx;
v3=xx*(u3'*u3)*xx;
s0=sqrt(diag(v0));          # Homoskedastic formula
s1=sqrt(diag(v1));          # White formula
s1a=sqrt(diag(v1a));        # Scaled White formula
s2=sqrt(diag(v2));          # Andrews formula
s3=sqrt(diag(v3));          # Horn-Horn-Duncan formula
```

4.17 Measures of Fit

As we described in the previous chapter, a commonly reported measure of regression fit is the regression R^2 defined as

$$R^2 = 1 - \frac{\sum_{i=1}^n \widehat{e}_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}_y^2}.$$

where $\widehat{\sigma}_y^2 = n^{-1}\sum_{i=1}^n (y_i - \overline{y})^2$. R^2 can be viewed as an estimator of the population parameter

$$\rho^2 = \frac{\text{var}\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)}{\text{var}(y_i)} = 1 - \frac{\sigma^2}{\sigma_y^2}.$$

However, $\widehat{\sigma}^2$ and $\widehat{\sigma}_y^2$ are biased estimators. Theil (1961) proposed replacing these by the unbiased versions s^2 and $\widetilde{\sigma}_y^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \overline{y})^2$ yielding what is known as **R-bar-squared** or **adjusted R-squared**:

$$\overline{R}^2 = 1 - \frac{s^2}{\widetilde{\sigma}_y^2} = 1 - \frac{(n-1)\sum_{i=1}^n \widehat{e}_i^2}{(n-k)\sum_{i=1}^n (y_i - \overline{y})^2}.$$

While $\overline{R}^2$ is an improvement on R^2, a much better improvement is

$$\widetilde{R}^2 = 1 - \frac{\sum_{i=1}^n \widetilde{e}_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = 1 - \frac{\widetilde{\sigma}^2}{\widehat{\sigma}_y^2}$$

where $\widetilde{e}_i$ are the prediction errors (3.44) and $\widetilde{\sigma}^2$ is the MSPE from (3.47). As described in Section (4.11), $\widetilde{\sigma}^2$ is a good estimator of the out-of-sample mean-squared forecast error, so $\widetilde{R}^2$ is a good estimator of the percentage of the forecast variance which is explained by the regression forecast. In this sense, $\widetilde{R}^2$ is a good measure of fit.

One problem with R^2, which is partially corrected by $\overline{R}^2$ and fully corrected by $\widetilde{R}^2$, is that R^2 necessarily increases when regressors are added to a regression model. This occurs because R^2 is a negative function of the sum of squared residuals which cannot increase when a regressor is added. In contrast, $\overline{R}^2$ and $\widetilde{R}^2$ are non-monotonic in the number of regressors. $\widetilde{R}^2$ can even be negative, which occurs when an estimated model predicts worse than a constant-only model.

In the statistical literature the MSPE $\widetilde{\sigma}^2$ is known as the **leave-one-out cross validation** criterion, and is popular for model comparison and selection, especially in high-dimensional (nonparametric) contexts. It is equivalent to use $\widetilde{R}^2$ or $\widetilde{\sigma}^2$ to compare and select models. Models with high $\widetilde{R}^2$ (or low $\widetilde{\sigma}^2$) are better models in terms of expected out of sample squared error. In contrast, R^2 cannot be used for model selection, as it necessarily increases when regressors are added to a regression model. $\overline{R}^2$ is also an inappropriate choice for model selection (it tends to select models with too many parameters), though a justification of this assertion requires a study of the theory of model selection. Unfortunately, $\overline{R}^2$ is routinely used by some economists, possibly as a hold-over from previous generations.

In summary, it is recommended to calculate and report $\widetilde{R}^2$ and/or $\widetilde{\sigma}^2$ in regression analysis, and omit R^2 and $\overline{R}^2$.

Henri Theil

Henri Theil (1924-2000) of the Netherlands invented $\overline{R}^2$ and two-stage least squares, both of which are routinely seen in applied econometrics. He also wrote an early influential advanced textbook on econometrics (Theil, 1971).

4.18 Empirical Example

We again return to our wage equation, but use a much larger sample of all individuals with at least 12 years of education. For regressors we include years of education, potential work experience, experience squared, and dummy variable indicators for the following: female, female union member,

male union member, married female[1], married male, formerly married female[2], formerly married male, Hispanic, black, American Indian, Asian, and mixed race[3] . The available sample is 46,943 so the parameter estimates are quite precise and reported in Table 4.1. For standard errors we use the unbiased Horn-Horn-Duncan formula.

Table 4.1 displays the parameter estimates in a standard tabular format. The table clearly states the estimation method (OLS), the dependent variable (log(Wage)), and the regressors are clearly labeled. Both parameter estimates and standard errors are reported for all coefficients. In addition to the coefficient estimates, the table also reports the estimated error standard deviation and the sample size. These are useful summary measures of fit which aid readers.

Table 4.1
OLS Estimates of Linear Equation for Log(Wage)

	$\widehat{\beta}$	$s(\widehat{\beta})$
Education	0.117	0.001
Experience	0.033	0.001
Experience^2/100	-0.056	0.002
Female	-0.098	0.011
Female Union Member	0.023	0.020
Male Union Member	0.095	0.020
Married Female	0.016	0.010
Married Male	0.211	0.010
Formerly Married Female	-0.006	0.012
Formerly Married Male	0.083	0.015
Hispanic	-0.108	0.008
Black	-0.096	0.008
American Indian	-0.137	0.027
Asian	-0.038	0.013
Mixed Race	-0.041	0.021
Intercept	0.909	0.021
$\widehat{\sigma}$	0.565	
Sample Size	46,943	

Note: Standard errors are heteroskedasticity-consistent (Horn-Horn-Duncan formula)

As a general rule, it is advisable to always report standard errors along with parameter estimates. This allows readers to assess the precision of the parameter estimates, and as we will discuss in later chapters, form confidence intervals and t-tests for individual coefficients if desired.

The results in Table 4.1 confirm our earlier findings that the return to a year of education is approximately 12%, the return to experience is concave, that single women earn approximately 10% less then single men, and blacks earn about 10% less than whites. In addition, we see that Hispanics earn about 11% less than whites, American Indians 14% less, and Asians and Mixed races about 4% less. We also see there are wage premiums for men who are members of a labor union (about 10%), married (about 21%) or formerly married (about 8%), but no similar premiums are apparent for women.

[1] Defining "married" as marital code 1, 2, or 3.
[2] Defining "formerly married" as marital code 4, 5, or 6.
[3] Race code 6 or higher.

4.19 Multicollinearity

If $\boldsymbol{X}'\boldsymbol{X}$ is singular, then $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ and $\widehat{\boldsymbol{\beta}}$ are not defined. This situation is called **strict multicollinearity**, as the columns of $\boldsymbol{X}$ are linearly dependent, i.e., there is some $\boldsymbol{\alpha} \neq \boldsymbol{0}$ such that $\boldsymbol{X}\boldsymbol{\alpha} = \boldsymbol{0}$. Most commonly, this arises when sets of regressors are included which are identically related. For example, if $\boldsymbol{X}$ includes both the logs of two prices and the log of the relative prices, $\log(p_1)$, $\log(p_2)$ and $\log(p_1/p_2)$, then $\boldsymbol{X}'\boldsymbol{X}$ will necessarily be singular. When this happens, the applied researcher quickly discovers the error as the statistical software will be unable to construct $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Since the error is discovered quickly, this is rarely a problem for applied econometric practice.

The more relevant situation is **near multicollinearity**, which is often called "multicollinearity" for brevity. This is the situation when the $\boldsymbol{X}'\boldsymbol{X}$ matrix is near singular, when the columns of $\boldsymbol{X}$ are close to linearly dependent. This definition is not precise, because we have not said what it means for a matrix to be "near singular". This is one difficulty with the definition and interpretation of multicollinearity.

One potential complication of near singularity of matrices is that the numerical reliability of the calculations may be reduced. In practice this is rarely an important concern, except when the number of regressors is very large.

A more relevant implication of near multicollinearity is that individual coefficient estimates will be imprecise. We can see this most simply in a homoskedastic linear regression model with two regressors

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i,$$

and

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In this case

$$\operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n\left(1-\rho^2\right)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

The correlation ρ indexes collinearity, since as ρ approaches 1 the matrix becomes singular. We can see the effect of collinearity on precision by observing that the variance of a coefficient estimate $\sigma^2\left[n\left(1-\rho^2\right)\right]^{-1}$ approaches infinity as ρ approaches 1. Thus the more "collinear" are the regressors, the worse the precision of the individual coefficient estimates.

What is happening is that when the regressors are highly dependent, it is statistically difficult to disentangle the impact of β_1 from that of β_2. As a consequence, the precision of individual estimates are reduced. The imprecision, however, will be reflected by large standard errors, so there is no distortion in inference.

Some earlier textbooks overemphasized a concern about multicollinearity. A very amusing parody of these texts appeared in Chapter 23.3 of Goldberger's *A Course in Econometrics* (1991), which is reprinted below. To understand his basic point, you should notice how the estimation variance $\sigma^2\left[n\left(1-\rho^2\right)\right]^{-1}$ depends equally and symmetrically on the correlation ρ and the sample size n.

Arthur S. Goldberger

Art Goldberger (1930-2009) was one of the most distinguished members of the Department of Economics at the University of Wisconsin. His PhD thesis developed an early macroeconometric forecasting model (known as the Klein-Goldberger model) but most of his career focused on microeconometric issues. He was the leading pioneer of what has been called the Wisconsin Tradition of empirical work – a combination of formal econometric theory with a careful critical analysis of empirical work. Goldberger wrote a series of highly regarded and influential graduate econometric textbooks, including *Econometric Theory* (1964), *Topics in Regression Analysis* (1968), and *A Course in Econometrics* (1991).

Micronumerosity

Arthur S. Goldberger

A Course in Econometrics (1991), Chapter 23.3

Econometrics texts devote many pages to the problem of multicollinearity in multiple regression, but they say little about the closely analogous problem of small sample size in estimating a univariate mean. Perhaps that imbalance is attributable to the lack of an exotic polysyllabic name for "small sample size." If so, we can remove that impediment by introducing the term *micronumerosity.*

Suppose an econometrician set out to write a chapter about small sample size in sampling from a univariate population. Judging from what is now written about multicollinearity, the chapter might look like this:

1. *Micronumerosity*

   The extreme case, "exact micronumerosity," arises when $n = 0$, in which case the sample estimate of μ is not unique. (Technically, there is a violation of the rank condition $n > 0$: the matrix 0 is singular.) The extreme case is easy enough to recognize. "Near micronumerosity" is more subtle, and yet very serious. It arises when the rank condition $n > 0$ is barely satisfied. Near micronumerosity is very prevalent in empirical economics.

2. *Consequences of micronumerosity*

   The consequences of micronumerosity are serious. Precision of estimation is reduced. There are two aspects of this reduction: estimates of μ may have large errors, and not only that, but $V_{\bar{y}}$ will be large.

   Investigators will sometimes be led to accept the hypothesis $\mu = 0$ because $\bar{y}/\widehat{\sigma}_{\bar{y}}$ is small, even though the true situation may be not that $\mu = 0$ but simply that the sample data have not enabled us to pick μ up.

   The estimate of μ will be very sensitive to sample data, and the addition of a few more observations can sometimes produce drastic shifts in the sample mean.

   The true μ may be sufficiently large for the null hypothesis $\mu = 0$ to be rejected, even though $V_{\bar{y}} = \sigma^2/n$ is large because of micronumerosity. But if the true μ is small (although nonzero) the hypothesis $\mu = 0$ may mistakenly be accepted.

3. *Testing for micronumerosity*

   Tests for the presence of micronumerosity require the judicious use of various fingers. Some researchers prefer a single finger, others use their toes, still others let their thumbs rule.

   A generally reliable guide may be obtained by counting the number of observations. Most of the time in econometric analysis, when n is close to zero, it is also far from infinity.

   Several test procedures develop critical values n^*, such that micronumerosity is a problem only if n is smaller than n^*. But those procedures are questionable.

4. *Remedies for micronumerosity*

   If micronumerosity proves serious in the sense that the estimate of μ has an unsatisfactorily low degree of precision, we are in the statistical position of not being able to make bricks without straw. The remedy lies essentially in the acquisition, if possible, of larger samples from the same population.

   But more data are no remedy for micronumerosity if the additional data are simply "more of the same." So obtaining lots of small samples from the same population will not help.

4.20 Clustered Sampling

In Section 4.2 we briefly mentioned clustered sampling as an alternative to the assumption of random sampling. We now introduce the framework in more detail and extend the primary results of this Chapter to encompass clustered dependence.

It might be easiest to understand the idea of clusters by considering a concrete example. Duflo, Dupas and Kremer (2011) investigate the impact of tracking (assigning students based on initial test score) on educational attainment in a randomized experiment. An extract of their data set is available on the textbook webpage in the file `DDK2011`.

In 2005, 140 primary schools in Kenya received funding to hire an extra first grade teacher to reduce class sizes. In half of the schools (selected randomly), students were assigned to classrooms based on an initial test score ("tracking"); in the remaining schools the students were randomly assigned to classrooms. For their analysis, the authors restricted attention to the 121 schools which initially had a single first-grade class, and if we further restrict attention to those with full data availability the resulting sample has 111 schools.

The key regression in the paper takes the form

$$TestScore_{ig} = -0.082 + 0.147 Tracking_g + e_{ig} \tag{4.45}$$

where $TestScore_{ig}$ is the standardized test score (normalized to have mean 0 and variance 1) of student i in school g, and $Tracking_g$ is a dummy equal to 1 if school g was tracking. The OLS estimates indicate that schools which tracked the students had an overall increase in test scores by 0.15 standard deviations, which is quite meaningful. More general versions of this regression are estimated, many of which take the form

$$TestScore_{ig} = \alpha + \gamma Tracking_g + \boldsymbol{x}_{ig}'\boldsymbol{\beta} + e_{ig} \tag{4.46}$$

where $\boldsymbol{x}_{ig}$ is a set of controls specific to the student (including age, sex and initial test score).

A difficulty with applying the classical regression framework is that student achievement is likely to be dependent within a given school. Student achievement may be affected by local demographics, individual teachers, and classmates, all of which imply dependence within a school. These concerns, however, do not suggest that achievement will be correlated across schools, so it seems reasonable to model achievement across schools as mutually independent.

In clustering contexts it is convenient to double index the observations as $(y_{ig}, \boldsymbol{x}_{ig})$ where $g = 1, ..., G$ indexes the cluster and $i = 1, ..., n_g$ indexes the individual within the g^{th} cluster. The number of observations per cluster n_g may vary across clusters. The number of clusters is G. The total number of observations is $n = \sum_{g=1}^{G} n_g$. In the Kenyan schooling example, the number of clusters (schools) in the estimation sample is $G = 111$, the number of students per school varies from 19 to 62, and the total number of observations is $n = 5269$.

While it is typical to write the observations using the double index notation $(y_{ig}, \boldsymbol{x}_{ig})$, it is also useful to use cluster-level notation. Let $\boldsymbol{y}_g = (y_{1g}, ..., y_{n_g g})'$ and $\boldsymbol{X}_g = (\boldsymbol{x}_{1g}, ..., \boldsymbol{x}_{n_g g})'$ denote the $n_g \times 1$ vector of dependent variables and $n_g \times k$ matrix of regressors for the g^{th} cluster. A linear regression model can be written for the individual observations as

$$y_{ig} = \boldsymbol{x}_{ig}'\boldsymbol{\beta} + e_{ig}$$

and using cluster notation as

$$\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{e}_g \tag{4.47}$$

where $\boldsymbol{e}_g = (e_{1g}, ..., e_{n_g g})'$ is a $n_g \times 1$ error vector.

Using this notation we can write the sums over the observations using the double sum $\sum_{g=1}^{G}\sum_{i=1}^{n_g}$. This is the sum across clusters of the sum across observations within each cluster. The OLS estimator can be written as

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} \boldsymbol{x}_{ig}\boldsymbol{x}_{ig}'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} \boldsymbol{x}_{ig}y_{ig}\right)$$

or

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{y}_g\right). \tag{4.48}$$

The OLS residuals are $\widehat{e}_{ig} = y_{ig} - \boldsymbol{x}_{ig}'\widehat{\boldsymbol{\beta}}$ in individual level notation and $\widehat{\boldsymbol{e}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g\widehat{\boldsymbol{\beta}}$ in cluster level notation.

The standard clustering assumption is that the clusters are known to the researcher and that the observations are independent across clusters.

Assumption 4.20.1 *The clusters $(\boldsymbol{y}_g, \boldsymbol{X}_g)$ are mutually independent across clusters g.*

In our example, clusters are schools. In other common applications, cluster dependence has been assumed within individual classrooms, families, villages, regions, and within larger units such as industries and states. This choice is up to the researcher, though the justification will depend on the context, the nature of the data, and will reflect information and assumptions on the dependence structure across observations.

The model is a linear regression under the assumption

$$\mathbb{E}\left(\boldsymbol{e}_g \mid \boldsymbol{X}_g\right) = 0. \tag{4.49}$$

This is the same as assuming that the individual errors are conditionally mean zero

$$\mathbb{E}\left(e_{ig} \mid \boldsymbol{X}_g\right) = 0$$

or that the conditional mean of $\boldsymbol{y}_g$ given $\boldsymbol{X}_g$ is linear. As in the independent case, equation (4.49) means that the linear regression model is correctly specified. In the clustered regression model, this requires that all all interaction effects within clusters have been accounted for in the specification of the individual regressors $\boldsymbol{x}_{ig}$.

In the regression (4.45), the conditional mean is necessarily linear and satisfies (4.49) since the regressor $Tracking_g$ is a dummy variable at the cluster level. In the regression (4.46) with individual controls, (4.49) requires that the achievement of any student is unaffected by the individual controls (e.g. age, sex and initial test score) of other students within the same school.

Given (4.49), we can calculate the mean of the OLS estimator. Substituting (4.47) into (4.48) we find

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right).$$

The mean of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ conditioning on all the regressors is

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid \boldsymbol{X}\right) &= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \mathbb{E}\left(\boldsymbol{e}_g \mid \boldsymbol{X}\right)\right) \\
&= \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \mathbb{E}\left(\boldsymbol{e}_g \mid \boldsymbol{X}_g\right)\right) \\
&= \mathbf{0}.
\end{aligned}$$

The first equality holds by linearity, the second by Assumption 4.20.1 and the third by (4.49).

This shows that OLS is unbiased under clustering if the conditional mean is linear.

Theorem 4.20.1 *In the clustered linear regression model (Assumption 4.20.1 and (4.49))*

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$$

Now consider the covariance matrix of $\widehat{\boldsymbol{\beta}}$. Let

$$\boldsymbol{\Sigma}_g = \mathbb{E}\left(\boldsymbol{e}_g \boldsymbol{e}_g' \mid \boldsymbol{X}_g\right)$$

denote the $n_g \times n_g$ conditional covariance matrix of the errors within the g^{th} cluster. Since the observations are independent across clusters,

$$\begin{aligned}
\operatorname{var}\left(\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right) \mid \boldsymbol{X}\right) &= \sum_{g=1}^{G} \operatorname{var}\left(\boldsymbol{X}_g' \boldsymbol{e}_g \mid \boldsymbol{X}_g\right) \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' \mathbb{E}\left(\boldsymbol{e}_g \boldsymbol{e}_g' \mid \boldsymbol{X}_g\right) \boldsymbol{X}_g \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{\Sigma}_g \boldsymbol{X}_g \\
&\stackrel{def}{=} \boldsymbol{\Omega}_n
\end{aligned} \tag{4.50}$$

It follows that

$$\begin{aligned} \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} &= \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{\Omega}_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \end{aligned} \tag{4.51}$$

where we write $\boldsymbol{X}'\boldsymbol{X} = \sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{X}_g = \sum_{g=1}^{G}\sum_{i=1}^{n_g} \boldsymbol{x}_{ig}\boldsymbol{x}_{ig}'$.

This differs from the formula in the independent case due to the correlation between observations within clusters. The magnitude of the difference depends on the degree of correlation between observations within clusters and the number of observations within clusters. To see this, suppose that all clusters have the same number of observations $n_g = N$, $\mathbb{E}\left(e_{ig}^2 \mid \boldsymbol{x}_g\right) = \sigma^2$, $\mathbb{E}\left(e_{ig}e_{\ell g} \mid \boldsymbol{x}_g\right) = \sigma^2\rho$ for $i \neq \ell$, and the regressors $\boldsymbol{x}_{ig}$ do not vary within a cluster. In this case the exact variance of the OLS estimator equals

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \sigma^2 \left(1 + \rho\left(N - 1\right)\right).$$

If $\rho > 0$, this shows that the actual variance is appropriately a multiple ρN of the conventional formula. In the Kenyan school example, the average cluster size is 48, so if the correlation between students is $\rho = 0.25$ the actual variance exceeds the conventional formula by a factor of about twelve. In this case the correct standard errors (the square root of the variance) should be a multiple of about three times the conventional formula. This is a substantial difference, and should not be neglected.

The typical solution is to use a covariance matrix estimate which extends the robust White formula to allow for general correlation within clusters. Recall that the insight of the White covariance estimator is that the squared error e_i^2 is unbiased for $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma_i^2$. Similarly with cluster dependence the matrix $\boldsymbol{e}_g\boldsymbol{e}_g'$ is unbiased for $\mathbb{E}\left(\boldsymbol{e}_g\boldsymbol{e}_g' \mid \boldsymbol{X}_g\right) = \boldsymbol{\Sigma}_g$. This means that an unbiased estimate for (4.50) is $\widetilde{\boldsymbol{\Omega}}_n = \sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{e}_g\boldsymbol{e}_g'\boldsymbol{X}_g$. This is not feasible, but we can replace the unknown errors by the OLS residuals to obtain the estimator

$$\begin{aligned} \widehat{\boldsymbol{\Omega}}_n &= \sum_{g=1}^{G} \boldsymbol{X}_g'\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{X}_g \\ &= \sum_{g=1}^{G}\sum_{i=1}^{n_g}\sum_{\ell=1}^{n_g} \boldsymbol{x}_{ig}\boldsymbol{x}_{\ell g}'\widehat{e}_{ig}\widehat{e}_{\ell g} \\ &= \sum_{g=1}^{G} \left(\sum_{i=1}^{n_g} \boldsymbol{x}_{ig}\widehat{e}_{ig}\right)\left(\sum_{\ell=1}^{n_g} \boldsymbol{x}_{\ell g}\widehat{e}_{\ell g}\right)'. \end{aligned} \tag{4.52}$$

The three expressions in (4.50) give three equivalent formula which could be used to calculate $\widehat{\boldsymbol{\Omega}}_n$. The final expression writes $\widehat{\boldsymbol{\Omega}}_n$ in terms of the cluster sums $\sum_{\ell=1}^{n_g} \boldsymbol{x}_{\ell g}\widehat{e}_{\ell g}$ which is basis for our example R and MATLAB codes shown below.

Given the expressions (4.50)-(4.51), a natural cluster covariance matrix estimator takes the form

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = a_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \widehat{\boldsymbol{\Omega}}_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{4.53}$$

where the term a_n is a possible finite-sample adjustment. The Stata cluster command uses

$$a_n = \left(\frac{n-1}{n-k}\right)\left(\frac{G}{G-1}\right). \tag{4.54}$$

The factor $G/(G-1)$ was derived by Chris Hansen (2007) in the context of equal-sized clusters to improve performance when the number of clusters G is small. The factor $(n-1)/(n-k)$ is an

ad hoc generalization which nests the adjustment used in (4.38), since when $G = n$ we have the simplification $a_n = n/(n-k)$.

Alternative cluster-robust covariance matrix estimators can be constructed using cluster-level prediction errors such as

$$\widetilde{e}_g = \boldsymbol{y}_g - \boldsymbol{X}_g \widehat{\boldsymbol{\beta}}_{-g}$$

where $\widehat{\boldsymbol{\beta}}_{-g}$ is the least-squares estimator omitting cluster g. We then have the robust covariance matrix estimator

$$\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \widetilde{e}_g \widetilde{e}_g' \boldsymbol{X}_g \right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Similarly to the heteroskedastic-robust case, you can show that $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is a conservative estimator for $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ in the sense that the conditional expectation of $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ exceeds $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$. This covariance matrix estimator is more cumbersome to implement, however, as the cluster-level prediction errors do not have a simple computational form so require a loop to estimate.

To illustrate in the context of the Kenyan schooling example, we present the regression of student test scores on the school-level tracking dummy, with two standard errors displayed. The first (in parenthesis) is the conventional robust standard error. The second [in square brackets] is the clustered standard error, where clustering is at the level of the school.

$$\begin{array}{rccc} TestScore_{ig} = - & 0.082 & + & 0.147 \quad Tracking_g + e_{ig} \\ & (0.020) & & (0.028) \\ & [0.054] & & [0.077] \end{array} \tag{4.55}$$

We can see that the cluster-robust standard errors are roughly three times the conventional robust standard errors. Consequently, confidence intervals for the coefficients are greatly affected by the choice.

For illustration, we list here the commands needed to produce the regression results with clustered standard errors in Stata, R, and MATLAB.

Stata do File

```
*       Load data:
use "DDK2011.dta"
*       Standard the test score variable to have mean zero and unit variance:
egen testscore = std(totalscore)
*       Regression with standard errors clustered at the school level:
reg testscore tracking, cluster(schoolid)
```

You can see that clustered standard errors are simple to calculate in Stata.

R Program File

```
# Load the data and create variables
data <- read.table("DDK2011.txt",header=TRUE,sep="\t")
y <- scale(as.matrix(data$totalscore))
n <- nrow(y)
x <- cbind(as.matrix(data$tracking),matrix(1,n,1))
schoolid <- as.matrix(data$schoolid)
k <- ncol(x)
invx <- solve(t(x)%*%x)
beta <- invx%*%t(x)%*%y
xe <- x*rep(y-x%*%beta,times=k)
# Clustered robust standard error
xe_sum <- rowsum(xe,schoolid)
G <- nrow(xe_sum)
omega <- t(xe_sum)%*%xe_sum
scale <- G/(G-1)*(n-1)/(n-k)
V_clustered = scale*invx%*%omega%*%invx
se_clustered <- sqrt(diag(V_clustered))
print(beta)
print(se_clustered)
```

Programming clustered standard errors in R is also relatively easy due to the convenient `rowsum` command, which sums variables within clusters.

MATLAB Program File

```
% Load the data and create variables
data = xlsread('DDK2011.xlsx');
schoolid = data(:,2);
tracking = data(:,7);
totalscore = data(:,62);
y = (totalscore - mean(totalscore))./std(totalscore);
x = [tracking,ones(size(y,1),1)];
[n,k] = size(x);
invx = inv(x'*x);
beta = invx*(x'*y);
e = y - x*beta;
% Clustered robust standard error
[schools,~,schoolidx] = unique(schoolid);
G = size(schools,1);
cluster_sums = zeros(G,k);
for j = 1:k
cluster_sums(:,j) = accumarray(schoolidx,x(:,j).*e);end
omega = cluster_sums'*cluster_sums;
scale = G/(G-1)*(n-1)/(n-k);
V_clustered = scale*invx*omega*invx;
se_clustered = sqrt(diag(V_clustered));
display(beta);
display(se_clustered);
```

Here we see that programming clustered standard errors in MATLAB is less convenient than the other packages, but still can be executed with just a few lines of code. This example uses the `accumarray` command, which is similar to the `rowsum` command in R, but only can be applied to vectors (hence the loop across the regressors) and works best if the clusterid variable are indices (which is why the original *schoolid* variable is transformed into indices in *schoolidx*. Application of these commands requires considerable case and attention.

4.21 Inference with Clustered Samples

In this section we give some cautionary remarks and general advice about cluster-robust inference in econometric practice. There has been remarkably little theoretical research about the properties of cluster-robust methods – until quite recently – so these remarks may become dated rather quickly.

In many respects cluster-robust inference should be viewed similarly to heteroskedaticity-robust inference, with where a "cluster" in the cluster-robust case is interpreted similarly to an "observation" in the heteroskedasticity-robust case. In particular, the effective sample size should be viewed as the number of clusters, not the "sample size" n. This is because the cluster-robust covariance matrix estimator effectively treats each cluster as a single observation, and estimates the covariance matrix based on the variation across cluster means. Hence if there are only $G = 50$ clusters, inference should be viewed as (at best) similar to heteroskedasticity-robust inference with $n = 50$ observations. This is a bit unsettling, for if the number of regressors is large (say $k = 20$), then the covariance matrix will be estimated quite imprecisely.

Furthermore, most cluster-robust theory (for example, the work of Chris Hansen (2007)) assumes that the clusters are homogeneous, including the assumption that the cluster sizes are all identical. This turns out to be a very important simplication. When this is violated – when, for example, cluster sizes are highly heterogeneous – this should be viewed as roughly equivalent to the heteroskedasticity-robust case with an extremely high degree of heteroskedasticity. If observations themselves are i.i.d. then cluster sums have variances which are proportional to the cluster sizes, so if the latter is heterogeneous so will be the variances of the cluster sums. This also has a large effect on finite sample inference. When clusters are heterogeneous then cluster-robust inference is similar to heteroskedasticity-robust inference with highly heteroskedastic observations.

Put together, if the number of clusters G is small and the number of observations per cluster is highly varied, then we should interpret inferential statements with a great degree of caution. Unfortunately, this is the norm. Many empirical studies on U.S. data cluster at the "state" level, meaning that there are 50 or 51 clusters (the District of Columbia is typically treated as a state). The number of observations vary considerably across states, since the populations are highly unequal. Thus when you read empirical papers with individual-level data but clustered at the "state" level you should be very cautious, and recognize that this is equivalent to inference with a small number of extremely heterogeneous observations.

A further complication occurs when we are interested in treatment, as in the tracking example given in the previous section. In many cases (including Duflo, Dupas and Kremer (2011)) the interest is in the effect of a specific treatment which is applied at the cluster level (in their case, treatment applies to schools). In many cases (not, however, Duflo, Dupas and Kremer (2011)), the number of treated clusters is small relative to the total number of clusters, in an extreme case there is just a single treated cluster. Based on the reasoning given above, these applications should be interpreted as equivalent to heteroskedasticity-robust inference with a sparse dummy variable, as discussed in Section 4.15. As discussed there, standard error estimates can be erroneously small. In the extreme of a single treated cluster (in the example, if only a single school was tracked) then if the regression is estimated using the pure dummy (no intercept) design, the estimated *tracking* coefficient will have a cluster standard error of 0. In general, reported standard errors will understate the imprecision of parameter estimates.

A practical question which arises in the context of cluster-robust inference is "At what level should we cluster?" In some examples you could cluster at a very fine level, such as families or classrooms, or at higher levels of aggregation, such as neighborhoods, schools, towns, counties, or states. What is the correct level at which to cluster? Rules of thumb have been advocated by practitioners, but at present there is little formal analysis to provide useful guidance. What do we know? If cluster dependence is ignored or imposed at too fine a level, then variance estimators will be biased and inference will be inaccurate. Typically this means that standard errors will be too small, giving rise to spurious indications of significance and precision. On the other hand when cluster-robust inference is based on higher levels of dependence, then the precision of the covariance matrix estimators will decrease, meaning that standard errors will be very imprecise estimates of the actual sampling uncertain. This means that there is a trade-off between bias and variance in the estimation of the covariance matrix by cluster-robust methods. It is not at all clear – based on current theory – what to do. I state this emphatically. We really do not know what is the "correct" level at which to do cluster-robust inference. This is a very interesting question and should certainly be explored by econometric research.

Exercises

Exercise 4.1 For some integer k, set $\mu_k = \mathbb{E}(y^k)$.

(a) Construct an estimator $\widehat{\mu}_k$ for μ_k.

(b) Show that $\widehat{\mu}_k$ is unbiased for μ_k.

(c) Calculate the variance of $\widehat{\mu}_k$, say $\text{var}(\widehat{\mu}_k)$. What assumption is needed for $\text{var}(\widehat{\mu}_k)$ to be finite?

(d) Propose an estimator of $\text{var}(\widehat{\mu}_k)$.

Exercise 4.2 Calculate $E((\overline{y} - \mu)^3)$, the skewness of $\overline{y}$. Under what condition is it zero?

Exercise 4.3 Explain the difference between $\overline{y}$ and μ. Explain the difference between $n^{-1}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'$ and $\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)$.

Exercise 4.4 True or False. If $y_i = x_i\beta + e_i$, $x_i \in \mathbb{R}$, $\mathbb{E}(e_i \mid x_i) = 0$, and $\widehat{e}_i$ is the OLS residual from the regression of y_i on x_i, then $\sum_{i=1}^{n} x_i^2\widehat{e}_i = 0$.

Exercise 4.5 Prove (4.17) and (4.18)

Exercise 4.6 Prove Theorem 4.8.1.

Exercise 4.7 Let $\widetilde{\boldsymbol{\beta}}$ be the GLS estimator (4.19) under the assumptions (4.15) and (4.16). Assume that $\boldsymbol{\Omega} = c^2\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma}$ known and c^2 unknown. Define the residual vector $\widetilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$, and an estimator for c^2

$$\widetilde{c}^2 = \frac{1}{n-k}\widetilde{\boldsymbol{e}}'\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{e}}.$$

(a) Show (4.20).

(b) Show (4.21).

(c) Prove that $\widetilde{\boldsymbol{e}} = \boldsymbol{M}_1\boldsymbol{e}$, where $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}$.

(d) Prove that $\boldsymbol{M}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_1 = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}$.

(e) Find $\mathbb{E}\left(\widetilde{c}^2 \mid \boldsymbol{X}\right)$.

(f) Is $\widetilde{c}^2$ a reasonable estimator for c^2?

Exercise 4.8 Let $(y_i, \boldsymbol{x}_i)$ be a random sample with $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$. Consider the **Weighted Least Squares** (WLS) estimator of $\boldsymbol{\beta}$

$$\widetilde{\boldsymbol{\beta}}_{\text{wls}} = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}\right)$$

where $\boldsymbol{W} = \text{diag}\left(w_1, ..., w_n\right)$ and $w_i = x_{ji}^{-2}$, where x_{ji} is one of the $\boldsymbol{x}_i$.

(a) In which contexts would $\widetilde{\boldsymbol{\beta}}_{\text{wls}}$ be a good estimator?

(b) Using your intuition, in which situations would you expect that $\widetilde{\boldsymbol{\beta}}_{\text{wls}}$ would perform better than OLS?

Exercise 4.9 Show (4.33) in the homoskedastic regression model.

Exercise 4.10 Prove (4.41).

Exercise 4.11 Show (4.42) in the homoskedastic regression model.

Exercise 4.12 Let $\mu = \mathbb{E}(y_i)$, $\sigma^2 = \mathbb{E}\left((y_i - \mu)^2\right)$ and $\mu_3 = \mathbb{E}\left((y_i - \mu)^3\right)$ and consider the sample mean $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$. Find $\mathbb{E}\left((\overline{y} - \mu)^3\right)$ as a function of μ, σ^2, μ_3 and n.

Exercise 4.13 Take the simple regression model $y_i = x_i\beta + e_i$, $x_i \in \mathbb{R}$, $\mathbb{E}(e_i \mid x_i) = 0$. Define $\sigma_i^2 = \mathbb{E}(e_i^2 \mid x_i)$ and $\mu_{3i} = \mathbb{E}(e_i^3 \mid x_i)$ and consider the OLS coefficient $\widehat{\beta}$. Find $\mathbb{E}\left(\left(\widehat{\beta} - \beta\right)^3 \mid \boldsymbol{X}\right)$.

Exercise 4.14 Take a regression model with i.i.d. observations (y_i, x_i) and scalar x_i

$$y_i = x_i\beta + e_i$$

$$\mathbb{E}(e_i \mid x_i) = 0$$

The parameter of interest is $\theta = \beta^2$. Consider the OLS estimates $\widehat{\beta}$ and $\widehat{\theta} = \widehat{\beta}^2$.

(a) Find $\mathbb{E}(\widehat{\theta}|\boldsymbol{X})$ using our knowledge of $\mathbb{E}(\widehat{\beta}|\boldsymbol{X})$ and $V_{\widehat{\beta}} = \text{var}(\widehat{\beta}|\boldsymbol{X})$. Is $\widehat{\theta}$ biased for θ?

(b) Suggest an (approximate) biased-corrected estimator $\widehat{\theta}^*$ using an estimate $\widehat{V}_{\widehat{\beta}}$ for $V_{\widehat{\beta}}$.

(c) For $\widehat{\theta}^*$ to be potentially unbiased, which estimate of $V_{\widehat{\beta}}$ is most appropriate?

Under which conditions is $\widehat{\theta}^*$ unbiased?

Exercise 4.15 Consider an iid sample $\{y_i, \boldsymbol{x}_i\}$ $i = 1, ..., n$ where $\boldsymbol{x}_i$ is $k \times 1$. Assume the linear conditional expectation model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i \mid \boldsymbol{x}_i) &= 0 \end{aligned}$$

Assume that $n^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k$ (orthonormal regressors). Consider the OLS estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.

(a) Find $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \text{var}(\widehat{\boldsymbol{\beta}})$

(b) In general, are $\widehat{\beta}_j$ and $\widehat{\beta}_\ell$ for $j \neq \ell$ correlated or uncorrelated?

(c) Find a sufficient condition so that $\widehat{\beta}_j$ and $\widehat{\beta}_\ell$ for $j \neq \ell$ are uncorrelated.

Exercise 4.16 Take the linear homoskedastic CEF

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i|\boldsymbol{x}_i) &= 0 \\ \mathbb{E}(e_i^2|\boldsymbol{x}_i) &= \sigma^2 \end{aligned} \tag{4.56}$$

and suppose that y_i^* is measured with error. Instead of y_i^*, we observe y_i which satisfies

$$y_i = y_i^* + u_i$$

where u_i is measurement error. Suppose that e_i and u_i are independent and

$$\begin{aligned} \mathbb{E}(u_i|\boldsymbol{x}_i) &= 0 \\ \mathbb{E}(u_i^2|\boldsymbol{x}_i) &= \sigma_u^2(\boldsymbol{x}_i) \end{aligned}$$

(a) Derive an equation for y_i as a function of $\boldsymbol{x}_i$. Be explicit to write the error term as a function of the structural errors e_i and u_i. What is the effect of this measurement error on the model (4.56)?

(b) Describe the effect of this measurement error on OLS estimation of $\boldsymbol{\beta}$ in the feasible regression of the observed y_i on $\boldsymbol{x}_i$.

(c) Describe the effect (if any) of this measurement error on appropriate standard error calculation for $\widehat{\boldsymbol{\beta}}$.

Exercise 4.17 Suppose that for a pair of observables (y_i, x_i) with $x_i > 0$ that an economic model implies

$$\mathbb{E}\left(y_i \mid x_i\right) = \left(\gamma + \theta x_i\right)^{1/2}. \tag{4.57}$$

A friend suggests that (given an iid sample) you estimate γ and θ by the linear regression of y_i^2 on x_i, that is, to estimate the equation

$$y_i^2 = \alpha + \beta x_i + e_i. \tag{4.58}$$

(a) Investigate your friend's suggestion. Define $u_i = y_i - \left(\gamma + \theta x_i\right)^{1/2}$. Show that $\mathbb{E}\left(u_i \mid x_i\right) = 0$ is implied by (4.57).

(b) Use $y_i = \left(\gamma + \theta x_i\right)^{1/2} + u_i$ to calculate $\mathbb{E}\left(y_i^2 \mid x_i\right)$. What does this tell you about the implied equation (4.58)?

(c) Can you recover either γ and/or θ from estimation of (4.58)? Are additional assumptions required?

(d) Is this a reasonable suggestion?

Exercise 4.18 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \\ \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) &= \sigma^2 \end{aligned}$$

where $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})$, with $\boldsymbol{x}_{1i}$ $k_1 \times 1$ and $\boldsymbol{x}_{2i}$ $k_2 \times 1$. Consider the short regression

$$y_i = \boldsymbol{x}_{1i}'\widehat{\boldsymbol{\beta}}_1 + \widehat{e}_i$$

and define the error variance estimator

$$s^2 = \frac{1}{n - k_1}\sum_{i=1}^{n} \widehat{e}_i^2.$$

Find $\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right)$

Exercise 4.19 Let $\boldsymbol{y}$ be $n \times 1$, $\boldsymbol{X}$ be $n \times k$, and $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{C}$ where $\boldsymbol{C}$ is $k \times k$ and full-rank. Let $\widehat{\boldsymbol{\beta}}$ be the least-squares estimator from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$, and let $\widehat{\boldsymbol{V}}$ be the estimate of its asymptotic covariance matrix. Let $\widehat{\boldsymbol{\beta}}^*$ and $\widehat{\boldsymbol{V}}^*$ be those from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}^*$. Derive an expression for $\widehat{\boldsymbol{V}}^*$ as a function of $\widehat{\boldsymbol{V}}$.

Exercise 4.20 Take the model

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) &= \boldsymbol{0} \\ \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right) &= \boldsymbol{\Omega} \end{aligned}$$

Assume for simplicity that $\boldsymbol{\Omega}$ is known. Consider the OLS and GLS estimators $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}\right)$ and $\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}\right)$. Compute the (conditional) covariance between $\widehat{\boldsymbol{\beta}}$ and $\tilde{\beta}$:

$$\mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mid \boldsymbol{X}\right)$$

Find the (conditional) covariance matrix for $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$:

$$\mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right)\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right)' \mid \boldsymbol{X}\right)$$

Exercise 4.21 The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \\ \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) &= \sigma_i^2 \\ \boldsymbol{\Omega} &= diag(\sigma_1^2, ..., \sigma_n^2). \end{aligned}$$

The parameter β is estimated both by OLS $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$ and GLS $\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}$. Let $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$ denote the residuals. Let $\widehat{R}^2 = 1 - \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}/(\boldsymbol{y}^{*\prime}\boldsymbol{y}^*)$ and $\widetilde{R}^2 = 1 - \widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}}/(\boldsymbol{y}^{*\prime}\boldsymbol{y}^*)$ denote the equation R^2 where $\boldsymbol{y}^* = \boldsymbol{y} - \overline{y}$. If the error e_i is truly heteroskedastic will $\widehat{R}^2$ or $\widetilde{R}^2$ be smaller?

Exercise 4.22 An economist friend tells you that the assumption that the observations $(y_i, \boldsymbol{x}_i)$ are i.i.d. implies that the regression $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ is homoskedastic. Do you agree with your friend? How would you explain your position?

Exercise 4.23 Take the linear regression model with $\mathbb{E}\left(\boldsymbol{y} \mid \boldsymbol{X}\right) = \boldsymbol{X}\beta$. Define the *ridge regression* estimator

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$

where $\lambda > 0$ is a fixed constant. Find $E\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$. Is $\widehat{\boldsymbol{\beta}}$ biased for $\boldsymbol{\beta}$?

Exercise 4.24 Continue the empirical analysis in Exercise 3.22.

(a) Calculate standard errors using the homoskedasticity formula and using the four covariance matrices from Section 4.13.

(b) Repeat in your second programming language. Are they identical?

Exercise 4.25 Continue the empirical analysis in Exercise 3.24. Calculate standard errors using the Horn-Horn-Duncan method. Repeat in your second programming language. Are they identical?

Exercise 4.26 Extend the empirical analysis reported in Section 4.20. Do a regression of standardized test score (*totalscore* normalized to have zero mean and variance 1) on tracking, age, sex, being assigned to the contract teacher, and student's percentile in the initial distribution. Calculate standard errors using both the conventional robust formula, and clustering based on the school.

(a) Compare the two sets of standard errors. Which standard error changes the most by clustering? Which changes the least?

(b) How does the coefficient on *tracking* change by inclusion of the individual controls (in comparison to the results from (4.55))?

Chapter 5

Normal Regression and Maximum Likelihood

5.1 Introduction

This chapter introduces the normal regression model and the method of maximum likelihood. The normal regression model is a special case of the linear regression model. It is important as normality allows precise distributional characterizations and sharp inferences. It also provides a baseline for comparison with alternative inference methods, such as asymptotic approximations and the bootstrap.

The method of maximum likelihood is a powerful statistical method for parametric models (such as the normal regression model) and is widely used in econometric practice.

5.2 The Normal Distribution

We say that a random variable X has the **standard normal distribution**, or **Gaussian**, written $X \sim \mathrm{N}(0,1)$, if it has the density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad -\infty < x < \infty. \tag{5.1}$$

The standard normal density is typically written with the symbol $\phi(x)$ and the corresponding distribution function by $\Phi(x)$. It is a valid density function by the following result.

Theorem 5.2.1

$$\int_0^\infty \exp\left(-x^2/2\right) dx = \sqrt{\frac{\pi}{2}}. \tag{5.2}$$

All moments of the normal distribution are finite. Since the density is symmetric about zero all odd moments are zero. By integration by parts, you can show (see Exercises 5.2 and 5.3) that $\mathbb{E}\left(X^2\right) = 1$ and $\mathbb{E}\left(X^4\right) = 3$. In fact, for any positive integer m,

$$\mathbb{E}\left(X^{2m}\right) = (2m-1)!! = (2m-1)\cdot(2m-3)\cdots 1.$$

The notation $k!! = k\cdot(k-2)\cdots 1$ is known as the **double factorial**. For example, $\mathbb{E}\left(X^6\right) = 15$, $\mathbb{E}\left(X^8\right) = 105$, and $\mathbb{E}\left(X^{10}\right) = 945$.

We say that X has a **univariate normal distribution**, written $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$, if it has the density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

The mean and variance of X are μ and σ^2, respectively.

We say that the k-vector $\boldsymbol{X}$ has a **multivariate normal distribution**, written $\boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, if it has the joint density

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2} \det\left(\boldsymbol{\Sigma}\right)^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}\right), \qquad \boldsymbol{x} \in \mathbb{R}^k.$$

The mean and covariance matrix of $\boldsymbol{X}$ are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. By setting $k = 1$ you can check that the multivariate normal simplifies to the univariate normal.

For technical purposes it is useful to know the form of the moment generating and characteristic functions.

Theorem 5.2.2 *If $\boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ then its moment generating funtion is*

$$M(\boldsymbol{t}) = \mathbb{E}\left(\exp\left(\boldsymbol{t}'\boldsymbol{X}\right)\right) = \exp\left(\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\right)$$

(see Exercise 5.8) and its characteristic function is

$$C(\boldsymbol{t}) = \mathbb{E}\left(\exp\left(\mathrm{i}\boldsymbol{t}'\boldsymbol{X}\right)\right) = \exp\left(\mathrm{i}\boldsymbol{\mu}'\boldsymbol{\lambda} - \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\right)$$

(see Exercise 5.9).

An important property of normal random vectors is that affine functions are also multivariate normal.

Theorem 5.2.3 *If $\boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ and $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$, then $\boldsymbol{Y} \sim \mathrm{N}\left(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}'\right)$.*

One simple implication of Theorem 5.2.3 is that if $\boldsymbol{X}$ is multivariate normal, then each component of $\boldsymbol{X}$ is univariate normal.

Another useful property of the multivariate normal distribution is that uncorrelatedness is the same as independence. That is, if a vector is multivariate normal, subsets of variables are independent if and only if they are uncorrelated.

Theorem 5.2.4 *If $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ is multivariate normal, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are uncorrelated if and only if they are independent.*

The normal distribution is frequently used for inference to calculate critical values and p-values. This involves evaluating the normal cdf $\Phi(x)$ and its inverse. Since the cdf $\Phi(x)$ is not available in closed form, statistical textbooks have traditionally provided tables for this purpose. Such tables are not used currently as now these calculations are embedded in statistical software. For convenience, we list the appropriate commands in MATLAB and R to compute the cumulative distribution function of commonly used statistical distributions.

Numerical Cumulative Distribution Function Calculation
To calculate $\Pr(X \leq x)$ for given x

	MATLAB	R	Stata
$\mathrm{N}(0,1)$	`normcdf(x)`	`pnorm(x)`	`normal(x)`
χ^2_r	`chi2cdf(x,r)`	`pchisq(x,r)`	`chi2(r,x)`
t_r	`tcdf(x,r)`	`pt(x,r)`	`1-ttail(r,x)`
$F_{r,k}$	`fcdf(x,r,k)`	`pf(x,r,k)`	`F(r,k,x)`
$\chi^2_r(d)$	`ncx2cdf(x,r,d)`	`pchisq(x,r,d)`	`nchi2(r,d,x)`
$F_{r,k}(d)$	`ncfcdf(x,r,k,d)`	`pf(x,r,k,d)`	`1-nFtail(r,k,d,x)`

Here we list the appropriate commands to compute the inverse probabilities (quantiles) of the same distributions.

Numerical Quantile Calculation
To calculate x which solves $p = \Pr(X \leq x)$ for given p

	MATLAB	R	Stata
$\mathrm{N}(0,1)$	`norminv(p)`	`qnorm(p)`	`invnormal(p)`
χ^2_r	`chi2inv(p,r)`	`qchisq(p,r)`	`invchi2(r,p)`
t_r	`tinv(p,r)`	`qt(p,r)`	`invttail(r,1-p)`
$F_{r,k}$	`finv(p,r,k)`	`qf(p,r,k)`	`invF(r,k,p)`
$\chi^2_r(d)$	`ncx2inv(p,r,d)`	`qchisq(p,r,d)`	`invnchi2(r,d,p)`
$F_{r,k}(d)$	`ncfinv(p,r,k,d)`	`qf(p,r,k,d)`	`invnFtail(r,k,d,1-p)`

5.3 Chi-Square Distribution

Many important distributions can be derived as transformation of multivariate normal random vectors, including the chi-square, the student t, and the F. In this section we introduce the chi-square distribution.

Let $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_r)$ be multivariate standard normal and define $Q = \boldsymbol{X}'\boldsymbol{X}$. The distribution of Q is called **chi-square** with r degrees of freedom, written as $Q \sim \chi^2_r$.

The mean and variance of $Q \sim \chi^2_r$ are r and $2r$, respectively. (See Exercise 5.10.)

The chi-square distribution function is frequently used for inference (critical values and p-values). In practice these calculations are performed numerically by statistical software, but for completeness we provide the density function.

Theorem 5.3.1 *The density of* χ^2_r *is*

$$f(x) = \frac{1}{2^{r/2}\Gamma\left(\frac{r}{2}\right)} x^{r/2-1} e^{-x/2}, \qquad x > 0 \tag{5.3}$$

where $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}du$ *is the gamma function (Section 5.18).*

For some theoretical applications, including the study of the power of statistical tests, it is useful to define a non-central version of the chi-square distribution. When $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{I}_r)$ is multivariate normal, we say that $Q = \boldsymbol{X}'\boldsymbol{X}$ has a **non-central chi-square** distribution, with r degrees of freedom and non-centrality parameter $\lambda = \boldsymbol{\mu}'\boldsymbol{\mu}$, and is written as $Q \sim \chi^2_r(\lambda)$. The non-central chi-square simplifies to the central (conventional) chi-square when $\lambda = 0$, so that $\chi^2_r(0) = \chi^2_r$.

Theorem 5.3.2 *The density of* $\chi^2_r(\lambda)$ *is*

$$f(x) = \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i f_{r+2i}(x), \qquad x > 0 \tag{5.4}$$

where $f_{r+2i}(x)$ *is the* χ^2_{r+2i} *density function (5.3).*

Interestingly, as can be seen from the formula (5.4), the distribution of $\chi^2_r(\lambda)$ only depends on the scalar non-centrality parameter λ, not the entire mean vector $\boldsymbol{\mu}$.

A useful fact about the central and non-central chi-square distributions is that they also can be derived from multivariate normal distributions with general covariance matrices.

Theorem 5.3.3 *If* $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{A})$ *with* $\boldsymbol{A} > 0$*,* $r \times r$*, then* $\boldsymbol{X}'\boldsymbol{A}^{-1}\boldsymbol{X} \sim \chi^2_r(\lambda)$*, where* $\lambda = \boldsymbol{\mu}'\boldsymbol{A}^{-1}\boldsymbol{\mu}$*.*

In particular, Theorem 5.3.3 applies to the central chi-squared distribution, so if $\boldsymbol{X} \sim \mathrm{N}(0, \boldsymbol{A})$ then $\boldsymbol{X}'\boldsymbol{A}^{-1}\boldsymbol{X} \sim \chi^2_r$.

5.4 Student t Distribution

Let $Z \sim \mathrm{N}(0,1)$ and $Q \sim \chi^2_r$ be independent, and define $T = Z/\sqrt{Q/r}$. The distribution of T is called the **student** t with r degrees of freedom, and is written $T \sim t_r$. Like the chi-square, the distribution only depends on the degree of freedom parameter r.

Theorem 5.4.1 *The density of* T *is*

$$f(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}, \qquad -\infty < x < \infty.$$

The density function of the student t is bell-shaped like the normal density function, but the t has thicker tails. The t distribution has the property that moments below r are finite, but absolute moments greater than or equal to r are infinite.

The student t can also be seen as a generalization of the standard normal, for the latter is obtained as the limiting case where r is taken to infinity.

Theorem 5.4.2 *Let $f_r(x)$ be the student t density. As $r \to \infty$, $f_r(x) \to \phi(x)$.*

Another special case of the student t distribution occurs when $r = 1$ and is known as the **Cauchy** distribution. The Cauchy density function is

$$f(x) = \frac{1}{\pi (1 + x^2)}, \qquad -\infty < x < \infty.$$

A Cauchy random variable $T = Z_1/Z_2$ can also be derived as the ratio of two independent $\mathrm{N}(0,1)$ variables. The Cauchy has the property that it has no finite integer moments.

William Gosset

William S. Gosset (1876-1937) of England is most famous for his derivation of the student's t distribution, published in the paper "The probable error of a mean" in 1908. At the time, Gosset worked at Guiness Brewery, which prohibited its employees from publishing in order to prevent the possible loss of trade secrets. To circumvent this barrier, Gosset published under the pseudonym "Student". Consequently, this famous distribution is known as the student t rather than Gosset's t!

5.5 F Distribution

Let $Q_m \sim \chi^2_m$ and $Q_r \sim \chi^2_r$ be independent. The distribution of $F = (Q_m/m) / (Q_r/r)$ is called the F distribution with degree of freedom parameters m and r, and we write $F \sim F_{m,r}$.

Theorem 5.5.1 *The density of F is*

$$f(x) = \frac{\left(\frac{m}{r}\right)^{m/2} x^{m/2-1} \Gamma\left(\frac{m+r}{2}\right)}{\Gamma\left(\frac{m}{2}\right) \Gamma\left(\frac{r}{2}\right) \left(1 + \frac{m}{r}x\right)^{(m+r)/2}}, \qquad x > 0.$$

If $m = 1$ then we can write $Q_1 = Z^2$ where $Z \sim N(0,1)$, and $F = Z^2/(Q_r/r) = \left(Z/\sqrt{Q_r/r}\right)^2 = T^2$, the square of a student t with r degree of freedom. Thus the F distribution with $m = 1$ is equal to the squared student t distribution. In this sense the F distribution is a generalization of the student t.

As a limiting case, as $r \to \infty$ the F distribution simplifies to $F \to Q_m/m$, a normalized χ^2_m. Thus the F distribution is also a generalization of the χ^2_m distribution.

Theorem 5.5.2 *Let $f_{m,r}(x)$ be the density of mF. As $r \to \infty$, $f_{m,r}(x) \to f_m(x)$, the density of χ^2_m.*

Similarly with the non-central chi-square we define the non-central F distribution. If $Q_m \sim \chi^2_m(\lambda)$ and $Q_r \sim \chi^2_r$ are independent, then $F = (Q_m/m)/(Q_r/r)$ is called a **non-central** F with degree of freedom parameters m and r and non-centrality parameter λ.

5.6 Joint Normality and Linear Regression

Suppose the variables $(y, \boldsymbol{x})$ are jointly normally distributed. Consider the best linear predictor of y given $\boldsymbol{x}$

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \alpha + e.$$

By the properties of the best linear predictor, $\mathbb{E}(\boldsymbol{x}e) = 0$ and $\mathbb{E}(e) = 0$, so $\boldsymbol{x}$ and e are uncorrelated. Since $(e, \boldsymbol{x})$ is an affine transformation of the normal vector $(y, \boldsymbol{x})$, it follows that $(e, \boldsymbol{x})$ is jointly normal (Theorem 5.2.3). Since $(e, \boldsymbol{x})$ is jointly normal and uncorrelated they are independent (Theorem 5.2.4). Independence implies that

$$\mathbb{E}(e \mid \boldsymbol{x}) = \mathbb{E}(e) = 0$$

and

$$\mathbb{E}\left(e^2 \mid \boldsymbol{x}\right) = \mathbb{E}\left(e^2\right) = \sigma^2$$

which are properties of a homoskedastic linear CEF.

We have shown that when $(y, \boldsymbol{x})$ are jointly normally distributed, they satisfy a normal linear CEF

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \alpha + e$$

where

$$e \sim \mathrm{N}(0, \sigma^2)$$

is independent of $\boldsymbol{x}$.

This is a classical motivation for the linear regression model.

5.7 Normal Regression Model

The normal regression model is the linear regression model with an independent normal error

$$\begin{aligned} y &= \boldsymbol{x}'\boldsymbol{\beta} + e \\ e &\sim \mathrm{N}(0, \sigma^2). \end{aligned} \tag{5.5}$$

As we learned in Section 5.6, the normal regression model holds when $(y, \boldsymbol{x})$ are jointly normally distributed. Normal regression, however, does not require joint normality. All that is required is that the conditional distribution of y given $\boldsymbol{x}$ is normal (the marginal distribution of $\boldsymbol{x}$ is unrestricted). In this sense the normal regression model is broader than joint normality. Notice that for notational convenience we have written (5.5) so that $\boldsymbol{x}$ contains the intercept.

Normal regression is a parametric model, where likelihood methods can be used for estimation, testing, and distribution theory. The **likelihood** is the name for the joint probability density of the data, evaluated at the observed sample, and viewed as a function of the parameters. The maximum likelihood estimator is the value which maximizes this likelihood function. Let us now derive the likelihood of the normal regression model.

First, observe that model (5.5) is equivalent to the statement that the conditional density of y given $\boldsymbol{x}$ takes the form

$$f\left(y \mid \boldsymbol{x}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2\right).$$

Under the assumption that the observations are mutually independent, this implies that the conditional density of $(y_1, ..., y_n)$ given $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ is

$$\begin{aligned} f\left(y_1, ..., y_n \mid \boldsymbol{x}_1, ..., \boldsymbol{x}_n\right) &= \prod_{i=1}^{n} f\left(y_i \mid \boldsymbol{x}_i\right) \\ &= \prod_{i=1}^{n} \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right) \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right) \\ &\overset{def}{=} L(\boldsymbol{\beta}, \sigma^2) \end{aligned}$$

and is called the **likelihood function**.

For convenience, it is typical to work with the natural logarithm

$$\begin{aligned} \log f\left(y_1, ..., y_n \mid \boldsymbol{x}_1, ..., \boldsymbol{x}_n\right) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\ &\overset{def}{=} \log L(\boldsymbol{\beta}, \sigma^2) \end{aligned} \tag{5.6}$$

which is called the **log-likelihood function**.

The **maximum likelihood estimator** (**MLE**) $(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}})$ is the value which maximizes the log-likelihood. (It is equivalent to maximize the likelihood or the log-likelihood. See Exercise 5.15.) We can write the maximization problem as

$$(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^k,\ \sigma^2>0}{\operatorname{argmax}} \log L(\boldsymbol{\beta}, \sigma^2). \tag{5.7}$$

In most applications of maximum likelihood, the MLE must be found by numerical methods. However, in the case of the normal regression model we can find an explicit expression for $\widehat{\boldsymbol{\beta}}_{\text{mle}}$ and $\widehat{\sigma}^2_{\text{mle}}$ as functions of the data.

The maximizers $(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}})$ of (5.7) jointly solve the first-order conditions (FOC)

$$0 = \frac{\partial}{\partial\boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma^2)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{mle}}, \sigma^2=\widehat{\sigma}^2_{\text{mle}}} = \frac{1}{\widehat{\sigma}^2_{\text{mle}}}\sum_{i=1}^{n}\boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{mle}}\right) \tag{5.8}$$

$$0 = \frac{\partial}{\partial\sigma^2} \log L(\boldsymbol{\beta}, \sigma^2)\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}_{\text{mle}}, \sigma^2=\widehat{\sigma}^2_{\text{mle}}} = -\frac{n}{2\widehat{\sigma}^2_{\text{mle}}} + \frac{1}{\widehat{\sigma}^4_{\text{mle}}}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{mle}}\right)^2. \tag{5.9}$$

The first FOC (5.8) is proportional to the first-order conditions for the least-squares minimization problem of Section 3.6. It follows that the MLE satisfies

$$\widehat{\boldsymbol{\beta}}_{\text{mle}} = \left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right) = \widehat{\boldsymbol{\beta}}_{\text{ols}}.$$

That is, the MLE for $\boldsymbol{\beta}$ is algebraically identical to the OLS estimator.

Solving the second FOC (5.9) for $\widehat{\sigma}^2_{\text{mle}}$ we find

$$\widehat{\sigma}^2_{\text{mle}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{mle}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{ols}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 = \widehat{\sigma}^2_{\text{ols}}.$$

Thus the MLE for σ^2 is identical to the OLS/moment estimator from (3.33).

Since the OLS estimate and MLE under normality are equivalent, $\widehat{\boldsymbol{\beta}}$ is described by some authors as the maximum likelihood estimator, and by other authors as the least-squares estimator. It is important to remember, however, that $\widehat{\boldsymbol{\beta}}$ is only the MLE when the error e has a known normal distribution, and not otherwise.

Plugging the estimators into (5.6) we obtain the maximized log-likelihood

$$\log L\left(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}}\right) = -\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2_{\text{mle}}\right) - \frac{n}{2}. \tag{5.10}$$

The log-likelihood is typically reported as a measure of fit.

It may seem surprising that the MLE $\widehat{\boldsymbol{\beta}}_{\text{mle}}$ is numerically equal to the OLS estimator, despite emerging from quite different motivations. It is not completely accidental. The least-squares estimator minimizes a particular sample loss function – the sum of squared error criterion – and most loss functions are equivalent to the likelihood of a specific parametric distribution, in this case the normal regression model. In this sense it is not surprising that the least-squares estimator can be motivated as either the minimizer of a sample loss function or as the maximizer of a likelihood function.

Carl Friedrich Gauss

The mathematician Carl Friedrich Gauss (1777-1855) proposed the normal regression model, and derived the least squares estimator as the maximum likelihood estimator for this model. He claimed to have discovered the method in 1795 at the age of eighteen, but did not publish the result until 1809. Interest in Gauss's approach was reinforced by Laplace's simultaneous discovery of the central limit theorem, which provided a justification for viewing random disturbances as approximately normal.

5.8 Distribution of OLS Coefficient Vector

In the normal linear regression model we can derive exact sampling distributions for the OLS/MLE estimates, residuals, and variance estimate. In this section we derive the distribution of the OLS coefficient estimate.

The normality assumption $e_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(0, \sigma^2\right)$ combined with independence of the observations has the multivariate implication

$$\boldsymbol{e} \mid \boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_n\sigma^2\right).$$

That is, the error vector $\boldsymbol{e}$ is independent of $\boldsymbol{X}$ and is normally distributed.

Recall that the OLS estimator satisfies

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}$$

which is a linear function of $\boldsymbol{e}$. Since linear functions of normals are also normal (Theorem 5.2.3), this implies that conditional on $\boldsymbol{X}$,

$$\begin{aligned}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\Big|_{\boldsymbol{X}} &\sim \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathrm{N}\left(0, \boldsymbol{I}_n\sigma^2\right) \\ &\sim \mathrm{N}\left(0, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) \\ &= \mathrm{N}\left(0, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).\end{aligned}$$

An alternative way of writing this is

$$\widehat{\boldsymbol{\beta}}\Big|_{\boldsymbol{X}} \sim \mathrm{N}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

This shows that under the assumption of normal errors, the OLS estimate has an exact normal distribution.

Theorem 5.8.1 *In the linear regression model,*

$$\widehat{\boldsymbol{\beta}}\Big|_{\boldsymbol{X}} \sim \mathrm{N}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

Theorems 5.2.3 and 5.8.1 imply that any affine function of the OLS estimate is also normally distributed, including individual estimates. Letting β_j and $\widehat{\beta}_j$ denote the j^{th} elements of $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$, we have

$$\widehat{\beta}_j\Big|_{\boldsymbol{X}} \sim \mathrm{N}\left(\beta_j, \sigma^2\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}\right). \tag{5.11}$$

5.9 Distribution of OLS Residual Vector

Now consider the OLS residual vector. Recall from (3.31) that $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. This shows that $\widehat{\boldsymbol{e}}$ is linear in $\boldsymbol{e}$. So conditional on $\boldsymbol{X}$,

$$\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}|_{\boldsymbol{X}} \sim \mathrm{N}\left(0, \sigma^2\boldsymbol{M}\boldsymbol{M}\right) = \mathrm{N}\left(0, \sigma^2\boldsymbol{M}\right)$$

the final equality since $\boldsymbol{M}$ is idempotent (see Section 3.12). This shows that the residual vector has an exact normal distribution.

Furthermore, it is useful to understand the joint distribution of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{e}}$. This is easiest done by writing the two as a stacked linear function of the error $\boldsymbol{e}$. Indeed,

$$\begin{pmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\boldsymbol{e}} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e} \\ \boldsymbol{M}\boldsymbol{e} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\ \boldsymbol{M} \end{pmatrix}\boldsymbol{e}$$

which is is a linear function of $\boldsymbol{e}$. The vector thus has a joint normal distribution with covariance matrix

$$\begin{pmatrix} \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} & 0 \\ 0 & \sigma^2\boldsymbol{M} \end{pmatrix}.$$

The covariance is zero because $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{M} = 0$ as $\boldsymbol{X}'\boldsymbol{M} = 0$ from (3.28). Since the covariance is zero, it follows that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{e}}$ are statistically independent (Theorem 5.2.4).

Theorem 5.9.1 *In the linear regression model,* $\widehat{\boldsymbol{e}}|_{\boldsymbol{X}} \sim \mathrm{N}\left(0, \sigma^2\boldsymbol{M}\right)$ *and is independent of* $\widehat{\boldsymbol{\beta}}$.

The fact that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{e}}$ are independent implies that $\widehat{\boldsymbol{\beta}}$ is independent of any function of the residual vector, including individual residuals $\widehat{e}_i$ and the variance estimate s^2 and $\widehat{\sigma}^2$.

5.10 Distribution of Variance Estimate

Next, consider the variance estimator s^2 from (4.30). Using (3.35), it satisfies $(n-k)\,s^2 = \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}} = \boldsymbol{e}'\boldsymbol{M}\,\boldsymbol{e}$. The spectral decomposition of $\boldsymbol{M}$ (see equation (A.10)) is $\boldsymbol{M} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$ and $\boldsymbol{\Lambda}$ is diagonal with the eigenvalues of $\boldsymbol{M}$ on the diagonal. Since $\boldsymbol{M}$ is idempotent with rank $n-k$ (see Section 3.12) it has $n-k$ eigenvalues equalling 1 and k eigenvalues equalling 0, so

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}_k \end{bmatrix}.$$

Let $\boldsymbol{u} = \boldsymbol{H}'\boldsymbol{e} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_n\sigma^2\right)$ (see Exercise 5.13) and partition $\boldsymbol{u} = (\boldsymbol{u}_1', \boldsymbol{u}_2')'$ where $\boldsymbol{u}_1 \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_{n-k}\sigma^2\right)$. Then

$$\begin{aligned}
(n-k)\,s^2 &= \boldsymbol{e}'\boldsymbol{M}\,\boldsymbol{e} \\
&= \boldsymbol{e}'\boldsymbol{H}\begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{H}'\boldsymbol{e} \\
&= \boldsymbol{u}'\begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{u} \\
&= \boldsymbol{u}_1'\boldsymbol{u}_1 \\
&\sim \sigma^2\chi^2_{n-k}.
\end{aligned}$$

We see that in the normal regression model the exact distribution of s^2 is a scaled chi-square.

Since $\widehat{\boldsymbol{e}}$ is independent of $\widehat{\boldsymbol{\beta}}$ it follows that s^2 is independent of $\widehat{\boldsymbol{\beta}}$ as well.

Theorem 5.10.1 *In the linear regression model,*

$$\frac{(n-k)\,s^2}{\sigma^2} \sim \chi^2_{n-k}$$

and is independent of $\widehat{\boldsymbol{\beta}}$.

5.11 t-statistic

An alternative way of writing (5.11) is

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}}} \sim \mathrm{N}(0,1).$$

This is sometimes called a **standardized** statistic, as the distribution is the standard normal.

Now take the standardized statistic and replace the unknown variance σ^2 with its estimate s^2. We call this a **t-ratio** or **t-statistic**

$$T = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{s^2\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}}} = \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)}$$

where $s(\widehat{\beta}_j)$ is the classical (homoskedastic) standard error for $\widehat{\beta}_j$ from (4.43). We will sometimes write the t-statistic as $T(\beta_j)$ to explicitly indicate its dependence on the parameter value β_j, and

sometimes will simplify notation and write the t-statistic as T when the dependence is clear from the context.

By some algebraic re-scaling we can write the t-statistic as the ratio of the standardized statistic and the square root of the scaled variance estimate. Since the distributions of these two components are normal and chi-square, respectively, and independent, then we can deduce that the t-statistic has the distribution

$$\begin{aligned} T &= \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\right]_{jj}}} \Bigg/ \sqrt{\frac{(n-k)s^2}{\sigma^2} \Big/ (n-k)} \\ &\sim \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-k} / (n-k)}} \\ &\sim t_{n-k} \end{aligned}$$

a student t distribution with $n-k$ degrees of freedom.

This derivation shows that the t-ratio has a sampling distribution which depends only on the quantity $n-k$. The distribution does not depend on any other features of the data. In this context, we say that the distribution of the t-ratio is **pivotal**, meaning that it does not depend on unknowns.

The trick behind this result is scaling the centered coefficient by its standard error, and recognizing that each depends on the unknown σ only through scale. Thus the ratio of the two does not depend on σ. This trick (scaling to eliminate dependence on unknowns) is known as **studentization**.

Theorem 5.11.1 *In the normal regression model, $T \sim t_{n-k}$.*

An important caveat about Theorem 5.11.1 is that it only applies to the t-statistic constructed with the homoskedastic (old-fashioned) standard error estimate. It does not apply to a t-statistic constructed with any of the robust standard error estimates. In fact, the robust t-statistics can have finite sample distributions which deviate considerably from t_{n-k} even when the regression errors are independent $N(0, \sigma^2)$. Thus the distributional result in Theorem 5.11.1, and the use of the t distribution in finite samples, should only be applied to classical t-statistics.

5.12 Confidence Intervals for Regression Coefficients

An OLS estimate $\widehat{\beta}$ is a **point estimate** for a coefficient β. A broader concept is a **set or interval estimate** which takes the form $\widehat{C} = [\widehat{L}, \widehat{U}]$. The goal of an interval estimate $\widehat{C}$ is to contain the true value, e.g. $\beta \in \widehat{C}$, with high probability.

The interval estimate $\widehat{C}$ is a function of the data and hence is random.

An interval estimate $\widehat{C}$ is called a $1-\alpha$ **confidence interval** when $\Pr(\beta \in \widehat{C}) = 1-\alpha$ for a selected value of α. The value $1-\alpha$ is called the **coverage probability**. Typical choices for the coverage probability $1-\alpha$ are 0.95 or 0.90.

The probability calculation $\Pr(\beta \in \widehat{C})$ is easily mis-interpreted as treating β as random and $\widehat{C}$ as fixed. (The probability that β is in $\widehat{C}$.) This is not the appropriate interpretation. Instead, the correct interpretation is that the probability $\Pr(\beta \in \widehat{C})$ treats the point β as fixed and the set $\widehat{C}$ as random. It is the probability that the random set $\widehat{C}$ covers (or contains) the fixed true coefficient β.

There is not a unique method to construct confidence intervals. For example, one simple (yet silly) interval is

$$\widehat{C} = \begin{cases} \mathbb{R} & \text{with probability } 1-\alpha \\ \left\{\widehat{\beta}\right\} & \text{with probability } \alpha \end{cases}.$$

If $\widehat{\beta}$ has a continuous distribution, then by construction $\Pr(\beta \in \widehat{C}) = 1-\alpha$, so this confidence interval has perfect coverage. However, $\widehat{C}$ is uninformative about $\widehat{\beta}$ and is therefore not useful.

Instead, a good choice for a confidence interval for the regression coefficient β is obtained by adding and subtracting from the estimate $\widehat{\beta}$ a fixed multiple of its standard error:

$$\widehat{C} = \left[\widehat{\beta} - c \cdot s(\widehat{\beta}),\quad \widehat{\beta} + c \cdot s(\widehat{\beta})\right] \tag{5.12}$$

where $c > 0$ is a pre-specified constant. This confidence interval is symmetric about the point estimate $\widehat{\beta}$, and its length is proportional to the standard error $s(\widehat{\beta})$.

Equivalently, $\widehat{C}$ is the set of parameter values for β such that the t-statistic $T(\beta)$ is smaller (in absolute value) than c, that is

$$\widehat{C} = \{\beta : |T(\beta)| \le c\} = \left\{\beta : -c \le \frac{\widehat{\beta}-\beta}{s(\widehat{\beta})} \le c\right\}.$$

The coverage probability of this confidence interval is

$$\begin{aligned} \Pr\left(\beta \in \widehat{C}\right) &= \Pr\left(|T(\beta)| \le c\right) \\ &= \Pr\left(-c \le T(\beta) \le c\right). \end{aligned} \tag{5.13}$$

Since the t-statistic $T(\beta)$ has the t_{n-k} distribution, (5.13) equals $F(c) - F(-c)$, where $F(u)$ is the student t distribution function with $n-k$ degrees of freedom. Since $F(-c) = 1 - F(c)$ (see Exercise 5.19) we can write (5.13) as

$$\Pr\left(\beta \in \widehat{C}\right) = 2F(c) - 1.$$

This is the **coverage probability** of the interval $\widehat{C}$, and only depends on the constant c.

As we mentioned before, a confidence interval has the coverage probability $1-\alpha$. This requires selecting the constant c so that $F(c) = 1-\alpha/2$. This holds if c equals the $1-\alpha/2$ quantile of the t_{n-k} distribution. As there is no closed form expression for these quantiles, we compute their values numerically. For example, by `tinv(1-alpha/2,n-k)` in MATLAB. With this choice the confidence interval (5.12) has exact coverage probability $1-\alpha$. By default, Stata reports 95% confidence intervals $\widehat{C}$ for each estimated regression coefficient using the same formula.

Theorem 5.12.1 *In the normal regression model, (5.12) with* $c = F^{-1}(1-\alpha/2)$ *has coverage probability* $\Pr\left(\beta \in \widehat{C}\right) = 1-\alpha$.

When the degree of freedom is large the distinction between the student t and the normal distribution is negligible. In particular, for $n-k \ge 61$ we have $c \le 2.00$ for a 95% interval. Using this value we obtain the most commonly used confidence interval in applied econometric practice:

$$\widehat{C} = \left[\widehat{\beta} - 2s(\widehat{\beta}),\quad \widehat{\beta} + 2s(\widehat{\beta})\right]. \tag{5.14}$$

This is a useful rule-of-thumb. This 95% confidence interval $\widehat{C}$ is simple to compute and can be easily calculated from coefficient estimates and standard errors.

Theorem 5.12.2 *In the normal regression model, if* $n-k \geq 61$ *then (5.14) has coverage probability* $\Pr\left(\beta \in \widehat{C}\right) \geq 0.95$.

Confidence intervals are a simple yet effective tool to assess estimation uncertainty. When reading a set of empirical results, look at the estimated coefficient estimates and the standard errors. For a parameter of interest, compute the confidence interval $\widehat{C}$ and consider the meaning of the spread of the suggested values. If the range of values in the confidence interval are too wide to learn about β, then do not jump to a conclusion about β based on the point estimate alone.

5.13 Confidence Intervals for Error Variance

We can also construct a confidence interval for the regression error variance σ^2 using the sampling distribution of s^2 from Theorem 5.10.1, which states that in the normal regression model

$$\frac{(n-k)\,s^2}{\sigma^2} \sim \chi^2_{n-k}. \tag{5.15}$$

Let $F(u)$ denote the χ^2_{n-k} distribution function, and for some α set $c_1 = F^{-1}(\alpha/2)$ and $c_2 = F^{-1}(1-\alpha/2)$ (the $\alpha/2$ and $1-\alpha/2$ quantiles of the χ^2_{n-k} distribution). Equation (5.15) implies that

$$\Pr\left(c_1 \leq \frac{(n-k)\,s^2}{\sigma^2} \leq c_2\right) = F(c_2) - F(c_1) = 1-\alpha.$$

Rewriting the inequalities we find

$$\Pr\left((n-k)\,s^2/c_2 \leq \sigma^2 \leq (n-k)\,s^2/c_1\right) = 1-\alpha.$$

This shows that an exact $1-\alpha$ confidence interval for σ^2 is

$$C = \left[\frac{(n-k)\,s^2}{c_2}, \quad \frac{(n-k)\,s^2}{c_1}\right]. \tag{5.16}$$

Theorem 5.13.1 *In the normal regression model, (5.16) has coverage probability* $\Pr\left(\sigma^2 \in C\right) = 1-\alpha$.

The confidence interval (5.16) for σ^2 is asymmetric about the point estimate s^2, due to the latter's asymmetric sampling distribution.

5.14 t Test

A typical goal in an econometric exercise is to assess whether or not coefficient β equals a specific value β_0. Often the specific value to be tested is $\beta_0 = 0$ but this is not essential. This is called **hypothesis testing**, a subject which will be explored in detail in Chapter 9. In this section and the following we give a short introduction specific to the normal regression model.

For simplicity write the coefficient to be tested as β. The **null hypothesis** is

$$\mathbb{H}_0 : \beta = \beta_0. \tag{5.17}$$

This states that the hypothesis is that the true value of the coefficient β equals the hypothesized value β_0.

The alternative hypothesis is the complement of $\mathbb{H}_0$, and is written as

$$\mathbb{H}_1 : \beta \neq \beta_0.$$

This states that the true value of β does not equal the hypothesized value.

We are interested in testing $\mathbb{H}_0$ against $\mathbb{H}_1$. The method is to design a statistic which is informative about $\mathbb{H}_1$. If the observed value of the statistic is consistent with random variation under the assumption that $\mathbb{H}_0$ is true, then we deduce that there is no evidence against $\mathbb{H}_0$ and consequently do not reject $\mathbb{H}_0$. However, if the statistic takes a value which is unlikely to occur under the assumption that $\mathbb{H}_0$ is true, then we deduce that there is evidence against $\mathbb{H}_0$, and consequently we reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$. The steps are to design a test statistic and characterize its sampling distribution under the assumption that $\mathbb{H}_0$ is true to control the probability of making a false rejection.

The standard statistic to test $\mathbb{H}_0$ against $\mathbb{H}_1$ is the absolute value of the t-statistic

$$|T| = \left| \frac{\widehat{\beta} - \beta_0}{s(\widehat{\beta})} \right|. \tag{5.18}$$

If $\mathbb{H}_0$ is true, then we expect $|T|$ to be small, but if $\mathbb{H}_1$ is true then we would expect $|T|$ to be large. Hence the standard rule is to reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ for large values of the t-statistic $|T|$, and otherwise fail to reject $\mathbb{H}_0$. Thus the hypothesis test takes the form

$$\text{Reject } \mathbb{H}_0 \text{ if } |T| > c.$$

The constant c which appears in the statement of the test is called the **critical value**. Its value is selected to control the probability of false rejections. When the null hypothesis is true, $|T|$ has an exact student t distribution (with $n-k$ degrees of freedom) in the normal regression model. Thus for a given value of c the probability of false rejection is

$$\begin{aligned} \Pr(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_0) &= \Pr(|T| > c \mid \mathbb{H}_0) \\ &= \Pr(T > c \mid \mathbb{H}_0) + \Pr(T < -c \mid \mathbb{H}_0) \\ &= 1 - F(c) + F(-c) \\ &= 2(1 - F(c)) \end{aligned}$$

where $F(u)$ is the t_{n-k} distribution function. This is the probability of false rejection, and is decreasing in the critical value c. We select the value c so that this probability equals a pre-selected value called the **significance level**, which is typically written as α. It is conventional to set $\alpha = 0.05$, though this is not a hard rule. We then select c so that $F(c) = 1 - \alpha/2$, which means that c is the $1 - \alpha/2$ quantile (inverse CDF) of the t_{n-k} distribution, the same as used for confidence intervals. With this choice, the decision rule "Reject $\mathbb{H}_0$ if $|T| > c$" has a significance level (false rejection probability) of α.

Theorem 5.14.1 *In the normal regression model, if the null hypothesis (5.17) is true, then for $|T|$ defined in (5.18), $|T| \sim t_{n-k}$. If c is set so that $\Pr(|t_{n-k}| \geq c) = \alpha$, then the test "Reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $|T| > c$" has significance level α.*

To report the result of a hypothesis test we need to pre-determine the significance level α in order to calculate the critical value c. This can be inconvenient and arbitrary. A simplification is to report what is known as the **p-value** of the test. In general, when a test takes the form "Reject $\mathbb{H}_0$ if $S > c$" and S has null distribution $G(u)$, then the p-value of the test is $p = 1 - G(S)$. A test with significance level α can be restated as "Reject $\mathbb{H}_0$ if $p < \alpha$". It is sufficient to report the p-value p, and we can interpret the value of p as indexing the test's strength of rejection of the null hypothesis. Thus a p-value of 0.07 might be interpreted as "nearly significant", 0.05 as "borderline significant", and 0.001 as "highly significant". In the context of the normal regression model, the p-value of a t-statistic $|T|$ is $p = 2(1 - F_{n-k}(|T|))$ where F_{n-k} is the CDF of the student t with $n - k$ degrees of freedom. For example, in MATLAB the calculation is `2*(1-tcdf(abs(t),n-k))`. In Stata, the default is that for any estimated regression, t-statistics for each estimated coefficient are reported along with their p-values calculated using this same formula. These t-statistics test the hypotheses that each coefficient is zero.

A p-value reports the stength of evidence against $\mathbb{H}_0$ but is not itself a probability. A common misunderstanding is that the p-value is the "probability that the null hypothesis is true". This is an incorrect interpretation. It is a statistic, and is random, and is a measure of the evidence against $\mathbb{H}_0$, nothing more.

5.15 Likelihood Ratio Test

In the previous section we described the t-test as the standard method to test a hypothesis on a single coefficient in a regression. In many contexts, however, we want to simultaneously assess a set of coefficients. In the normal regression model, this can be done by an F test, which can be derived from the likelihood ratio test.

Partition the regressors as $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}')$ and similarly partition the coefficient vector as $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. Then the regression model can be written as

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i. \tag{5.19}$$

Let $k = \dim(\boldsymbol{x}_i)$, $k_1 = \dim(\boldsymbol{x}_{1i})$, and $q = \dim(\boldsymbol{x}_{2i})$, so that $k = k_1 + q$. Partition the variables so that the hypothesis is that the second set of coefficients are zero, or

$$\mathbb{H}_0 : \boldsymbol{\beta}_2 = 0. \tag{5.20}$$

If $\mathbb{H}_0$ is true, then the regressors $\boldsymbol{x}_{2i}$ can be omitted from the regression. In this case we can write (5.19) as

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_i. \tag{5.21}$$

We call (5.21) the null model. The alternative hypothesis is that at least one element of $\boldsymbol{\beta}_2$ is non-zero and is written as

$$\mathbb{H}_1 : \boldsymbol{\beta}_2 \neq 0.$$

When models are estimated by maximum likelihood, a well-accepted testing procedure is to reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ for large values of the Likelihood Ratio – the ratio of the maximized likelihood function under $\mathbb{H}_1$ and $\mathbb{H}_0$, respectively. We now construct this statistic in the normal regression model. Recall from (5.10) that the maximized log-likelihood equals

$$\log L(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) = -\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2\right) - \frac{n}{2}.$$

We similarly need to calculate the maximized log-likelihood for the constrained model (5.21). By the same steps for derivation of the unconstrained MLE, we can find that the MLE for (5.21) is OLS of y_i on $\boldsymbol{x}_{1i}$. We can write this estimator as

$$\widetilde{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$$

with residual

$$\widetilde{e}_i = y_i - \boldsymbol{x}_{1i}'\widetilde{\boldsymbol{\beta}}_1$$

and error variance estimate

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widetilde{e}_i^2.$$

We use the tildes "˜" rather than the hats "^" above the constrained estimates to distinguish them from the unconstrained estimates. You can calculate similar to (5.10) that the maximized constrained log-likelihood is

$$\log L(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\sigma}^2) = -\frac{n}{2}\log\left(2\pi\widetilde{\sigma}^2\right) - \frac{n}{2}.$$

A classic testing procedure is to reject $\mathbb{H}_0$ for large values of the ratio of the maximized likelihoods. Equivalently, the test rejects $\mathbb{H}_0$ for large values of twice the difference in the log-likelihood functions. (Multiplying the likelihood difference by two turns out to be a useful scaling.) This equals

$$\begin{aligned} LR &= 2\left(\left(-\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2\right) - \frac{n}{2}\right) - \left(-\frac{n}{2}\log\left(2\pi\widetilde{\sigma}^2\right) - \frac{n}{2}\right)\right) \\ &= n\log\left(\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2}\right). \end{aligned} \tag{5.22}$$

The likelihood ratio test rejects for large values of LR, or equivalently (see Exercise 5.21), for large values of

$$F = \frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)/q}{\widehat{\sigma}^2/(n-k)}. \tag{5.23}$$

This is known as the F statistic for the test of hypothesis $\mathbb{H}_0$ against $\mathbb{H}_1$.

To develop an appropriate critical value, we need the null distribution of F. Recall from (3.35) that $n\widehat{\sigma}^2 = \boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P}$ with $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. Similarly, under $\mathbb{H}_0$, $n\widetilde{\sigma}^2 = \boldsymbol{e}'\boldsymbol{M}_1\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P}_1$ with $\boldsymbol{P}_1 = \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$. You can calculate that $\boldsymbol{M}_1 - \boldsymbol{M} = \boldsymbol{P} - \boldsymbol{P}_1$ is idempotent with rank q. Furthermore, $\left(\boldsymbol{M}_1 - \boldsymbol{M}\right)\boldsymbol{M} = 0$. It follows that $\boldsymbol{e}'\left(\boldsymbol{M}_1 - \boldsymbol{M}\right)\boldsymbol{e} \sim \chi^2_q$ and is independent of $\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}$. Hence

$$F = \frac{\boldsymbol{e}'\left(\boldsymbol{M}_1 - \boldsymbol{M}\right)\boldsymbol{e}/q}{\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}/(n-k)} \sim \frac{\chi^2_q/q}{\chi^2_{n-k}/(n-k)} \sim F_{q,n-k}$$

an exact F distribution with degrees of freedom q and $n-k$, respectively. Thus under $\mathbb{H}_0$, the F statistic has an exact F distribution.

The critical values are selected from the upper tail of the F distribution. For a given significance level α (typically $\alpha = 0.05$) we select the critical value c so that $\Pr\left(F_{q,n-k} \geq c\right) = \alpha$. (For example, in MATLAB the expression is `finv(1-α,q,n-k)`.) The test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $F > c$ and does not reject $\mathbb{H}_0$ otherwise. The p-value of the test is $p = 1 - G_{q,n-k}\left(F\right)$ where $G_{q,n-k}\left(u\right)$ is the $F_{q,n-k}$ distribution function. (In MATLAB, the p-value is computed as `1-fcdf(f,q,n-k)`.) It is equivalent to reject $\mathbb{H}_0$ if $F > c$ or $p < \alpha$.

In Stata, the command to test multiple coefficients takes the form 'test X1 X1' where X1 and X2 are the names of the variables whose coefficients are tested. Stata then reports the F statistic for the hypothesis that the coefficients are jointly zero along with the p-value calculated using the F distribution.

Theorem 5.15.1 *In the normal regression model, if the null hypothesis (5.20) is true, then for F defined in (5.23), $F \sim F_{q,n-k}$. If c is set so that $\Pr\left(F_{q,n-k} \geq c\right) = \alpha$, then the test "Reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $F > c$" has significance level α.*

Theorem 5.15.1 justifies the F test in the normal regression model with critical values taken from the F distribution.

5.16 Likelihood Properties

In this section we present some general properties of the likelihood which hold broadly – not just in normal regression.

Suppose that a random vector $\boldsymbol{y}$ has the conditional density $f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ where the function f is known, and the parameter vector $\boldsymbol{\theta}$ takes values in a parameter space $\boldsymbol{\Theta}$. The log-likelihood function for a random sample $\{\boldsymbol{y}_i \mid \boldsymbol{x}_i : i = 1, ..., n\}$ takes the form

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}).$$

A key property is that the expected log-likelihood is maximized at the true value of the parameter vector. At this point it is useful to make a notational distinction between a generic parameter value $\boldsymbol{\theta}$ and its true value $\boldsymbol{\theta}_0$. Set $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$.

Theorem 5.16.1 $\boldsymbol{\theta}_0 = \operatorname{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}(\log L(\boldsymbol{\theta}) \mid \boldsymbol{X})$

This motivates estimating $\boldsymbol{\theta}$ by finding the value which maximizes the log-likelihood function. This is the maximum likelihood estimator (MLE):

$$\widehat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \log L(\boldsymbol{\theta}).$$

The **score** of the likelihood function is the vector of partial derivatives with respect to the parameters, evaluated at the true values,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}.$$

The covariance matrix of the score is known as the **Fisher information**:

$$\boldsymbol{\mathcal{I}} = \operatorname{var}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0) \mid \boldsymbol{X}\right).$$

Some important properties of the score and information are now presented.

Theorem 5.16.2 *If* $\log f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})$ *is second differentiable and the support of* $\boldsymbol{y}$ *does not depend on* $\boldsymbol{\theta}$ *then*

1. $\mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \mid \boldsymbol{X}\right) = 0$
2. $\boldsymbol{\mathcal{I}} = \sum_{i=1}^{n} \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)' \mid \boldsymbol{x}_i\right)$
   $= -\mathbb{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log L(\boldsymbol{\theta}_0) \mid \boldsymbol{X}\right)$

The first result says that the score is mean zero. The second result shows that the variance of the score equals the negative expectation of the second derivative matrix. This is known as the **Information Matrix Equality**.

We now establish the famous Cramér-Rao Lower Bound.

Theorem 5.16.3 *(Cramér-Rao) Under the assumptions of Theorem 5.16.2, if* $\widetilde{\boldsymbol{\theta}}$ *is an unbiased estimator of* $\boldsymbol{\theta}$*, then* $\operatorname{var}\left(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{X}\right) \geq \boldsymbol{\mathcal{I}}^{-1}$.

Theorem 5.16.3 shows that the inverse of the information matrix is a lower bound for the covariance matrix of unbiased estimators. This result is similar to the Gauss-Markov Theorem which established a lower bound for unbiased estimators in homoskedastic linear regression.

Ronald Fisher

The British statistician Ronald Fisher (1890-1962) is one of the core founders of modern statistical theory. His contributions include the F distribution, p-values, the concept of Fisher information, and that of sufficient statistics.

5.17 Information Bound for Normal Regression

Recall the normal regression log-likelihood which has the parameters $\boldsymbol{\beta}$ and σ^2. The likelihood scores for this model are

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}\right) \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \boldsymbol{x}_i e_i\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}\right)^2 \\ &= \frac{1}{2\sigma^4} \sum_{i=1}^{n} \left(e_i^2 - \sigma^2\right).\end{aligned}$$

It follows that the information matrix is

$$\boldsymbol{\mathcal{I}} = \operatorname{var}\left(\begin{array}{c} \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \sigma^2) \\ \frac{\partial}{\partial \sigma^2} \log L(\boldsymbol{\beta}, \sigma^2) \end{array} \mid \boldsymbol{X}\right) = \begin{pmatrix} \frac{1}{\sigma^2} \boldsymbol{X}'\boldsymbol{X} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \tag{5.24}$$

(see Exercise 5.22). The Cramér-Rao Lower Bound is

$$\boldsymbol{\mathcal{I}}^{-1} = \begin{pmatrix} \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

This shows that the lower bound for estimation of $\boldsymbol{\beta}$ is $\sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$ and the lower bound for σ^2 is $2\sigma^4/n$.

Since in the homoskedastic linear regression model the OLS estimator is unbiased and has variance $\sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}$, it follows that OLS is Cramér-Rao efficient in the normal regression model, in the sense that no unbiased estimator has a lower variance matrix. This expands on the Gauss-Markov theorem, which stated that no linear unbiased estimator has a lower variance matrix in the homoskedastic regression model. Notice that that the results are complementary. Gauss-Markov efficiency concerns a more narrow class of estimators (linear) but allows a broader model class (linear homoskedastic rather than normal regression). The Cramér-Rao efficiency result is more powerful in that it does not restrict the class of estimators (beyond unbiasedness) but is more restrictive in the class of models allowed (normal regression).

In contrast, the unbiased estimator s^2 of σ^2 has variance $2\sigma^4/(n-k)$ (see Exercise 5.23) which is larger than the Cramér-Rao lower bound $2\sigma^4/n$. Thus in contrast to the coefficient estimator, the variance estimator is not Cramér-Rao efficient.

5.18 Gamma Function*

The normal and related distributions make frequent use of the what is known as the **gamma function**. For $\alpha > 0$ it is defined as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} \exp(-x)\, dx. \tag{5.25}$$

While it appears quite simple, it has some advanced properties. One is that $\Gamma(\alpha)$ does not have a close-form solution (except for special values of α). Thus it is typically represented using the symbol $\Gamma(\alpha)$ and implemented computationally using numerical methods.

Special values include

$$\Gamma(1) = \int_0^\infty \exp(-x)\, dx = 1 \tag{5.26}$$

and

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{5.27}$$

The latter holds by making the change of variables $x = u^2$ in (5.25) and applying (5.2).

By integration by parts you can show that it satisfies the property

$$\Gamma(1+\alpha) = \Gamma(\alpha)\alpha.$$

Combined with (5.26) we find that for positive integers n,

$$\Gamma(n) = (n-1)!$$

This shows that the gamma function is a continuous version of the factorial.

A useful fact is

$$\int_0^\infty y^{a-1} \exp(-by)\, dy = b^{-a}\Gamma(a) \tag{5.28}$$

which can be found by applying change-of-variables to the definition (5.25).

Another useful fact is for for $\alpha \in \mathbb{R}$

$$\lim_{n\to\infty} \frac{\Gamma(n+\alpha)}{\Gamma(n)\, n^\alpha} = 1. \tag{5.29}$$

5.19 Technical Proofs*

Proof of Theorem 5.2.1. Squaring expression (5.2)

$$\begin{aligned}\left(\int_0^\infty \exp\left(-x^2/2\right) dx\right)^2 &= \int_0^\infty \exp\left(-x^2/2\right) dx \int_0^\infty \exp\left(-u^2/2\right) du \\ &= \int_0^\infty \int_0^\infty \exp\left(-\left(x^2+u^2\right)/2\right) dxdu \\ &= \int_0^\infty \int_0^{\pi/2} r\exp\left(-r^2/2\right) d\theta dr \\ &= \frac{\pi}{2}\int_0^\infty r\exp\left(-r^2/2\right) dr \\ &= \frac{\pi}{2}.\end{aligned}$$

The third equality is the key. It makes the change-of-variables to polar coordinates $x = r\cos\theta$ and $u = r\sin\theta$ so that $x^2+u^2 = r^2$. The Jacobian of this transformation is r. The region of integration in the (x,u) units is the positive orthont (upper-right region), which corresponds to integrating θ from 0 to $\pi/2$ in polar coordinates. The final two equalities are simple integration. Taking the square root we obtain (5.2). ■

Proof of Theorem 5.2.3. Let $M_x(\boldsymbol{t}) = \exp\left(\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\right)$ be the moment generating function of $\boldsymbol{X}$ by Theorem 5.2.2. Then the MGF of $\boldsymbol{Y}$ is

$$\begin{aligned}\mathbb{E}\left(\exp\left(\boldsymbol{s}'\boldsymbol{Y}\right)\right) &= \mathbb{E}\exp\left(\boldsymbol{s}'\left(\boldsymbol{a}+\boldsymbol{B}\boldsymbol{X}\right)\right) \\ &= \exp\left(\boldsymbol{s}'\boldsymbol{a}\right)\mathbb{E}\exp\left(\boldsymbol{s}'\boldsymbol{B}\boldsymbol{X}\right) \\ &= \exp\left(\boldsymbol{s}'\boldsymbol{a}\right)M_x(\boldsymbol{B}'\boldsymbol{s}) \\ &= \exp\left(\boldsymbol{s}'\boldsymbol{a}\right)\exp\left(\boldsymbol{s}'\boldsymbol{B}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{s}'\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}'\boldsymbol{s}\right) \\ &= \exp\left(\boldsymbol{s}'\left(\boldsymbol{a}+\boldsymbol{B}\boldsymbol{\mu}\right) + \frac{1}{2}\boldsymbol{s}'\left(\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}'\right)\boldsymbol{s}\right)\end{aligned}$$

which is the MGF of $\mathrm{N}\left(\boldsymbol{a}+\boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}'\right)$. Thus $\boldsymbol{Y} \sim \mathrm{N}\left(\boldsymbol{a}+\boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}'\right)$ as claimed. ■

Proof of Theorem 5.2.4. Let k_1 and k_2 denote the dimensions of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and set $k = k_1 + k_2$. If the components are uncorrelated then the covariance matrix for $\boldsymbol{X}$ takes the form

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 & 0 \\ 0 & \boldsymbol{\Sigma}_2 \end{bmatrix}$$

In this case the joint density function of $\boldsymbol{X}$ equals

$$\begin{aligned}f(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \frac{1}{(2\pi)^{k/2}\left(\det\left(\boldsymbol{\Sigma}_1\right)\det\left(\boldsymbol{\Sigma}_2\right)\right)^{1/2}} \\ &\quad\cdot\exp\left(-\frac{\left(\boldsymbol{x}_1-\boldsymbol{\mu}_1\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{x}_1-\boldsymbol{\mu}_1\right) + \left(\boldsymbol{x}_2-\boldsymbol{\mu}_2\right)'\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{x}_2-\boldsymbol{\mu}_2\right)}{2}\right) \\ &= \frac{1}{(2\pi)^{k_1/2}\left(\det\left(\boldsymbol{\Sigma}_1\right)\right)^{1/2}}\exp\left(-\frac{\left(\boldsymbol{x}_1-\boldsymbol{\mu}_1\right)'\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{x}_1-\boldsymbol{\mu}_1\right)}{2}\right) \\ &\quad\cdot\frac{1}{(2\pi)^{k_2/2}\left(\det\left(\boldsymbol{\Sigma}_2\right)\right)^{1/2}}\exp\left(-\frac{\left(\boldsymbol{x}_2-\boldsymbol{\mu}_2\right)'\boldsymbol{\Sigma}_2^{-1}\left(\boldsymbol{x}_2-\boldsymbol{\mu}_2\right)}{2}\right).\end{aligned}$$

This is the product of two multivariate normal densities in $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Joint densities factor if (and only if) the components are independent. This shows that uncorrelatedness implies independence.

The converse (that independence implies uncorrelatedness) holds generally. ■

Proof of Theorem 5.3.1. We demonstrate that $Q = \boldsymbol{X}'\boldsymbol{X}$ has density function (5.3) by verifying that both have the same moment generating function (MGF). First, the MGF of $\boldsymbol{X}'\boldsymbol{X}$ is

$$\begin{aligned}
\mathbb{E}\left(\exp\left(t\boldsymbol{X}'\boldsymbol{X}\right)\right) &= \int_{-\infty}^{\infty} \exp\left(t\boldsymbol{x}'\boldsymbol{x}\right) \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{\boldsymbol{x}'\boldsymbol{x}}{2}\right) d\boldsymbol{x} \\
&= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{\boldsymbol{x}'\boldsymbol{x}}{2}(1-2t)\right) d\boldsymbol{x} \\
&= (1-2t)^{-r/2} \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{\boldsymbol{u}'\boldsymbol{u}}{2}\right) d\boldsymbol{u} \\
&= (1-2t)^{-r/2}. \qquad (5.30)
\end{aligned}$$

The fourth equality uses the change of variables $\boldsymbol{u} = (1-2t)^{1/2}\boldsymbol{x}$ and the final equality is the normal probability integral. Second, the MGF of the density (5.3) is

$$\begin{aligned}
\int_0^{\infty} \exp(tq) f(q) dq &= \int_0^{\infty} \exp(tq) \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{r/2}} q^{r/2-1} \exp(-q/2)\, dq \\
&= \int_0^{\infty} \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{r/2}} q^{r/2-1} \exp\left(-q\left(1/2-t\right)\right) dq \\
&= \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{r/2}} \left(1/2-t\right)^{-r/2} \Gamma\left(\frac{r}{2}\right) \\
&= (1-2t)^{-r/2}, \qquad (5.31)
\end{aligned}$$

the third equality using the gamma integral (5.28). The MGFs (5.30) and (5.31) are equal, verifying that (5.3) is the density of Q as claimed. ■

Proof of Theorem 5.3.2. As in the proof of Theorem 5.3.1 we verify that the MGF of $Q = \boldsymbol{X}'\boldsymbol{X}$ when $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{I}_r)$ is equal to the MGF of the density function (5.4).

First, we calculate the MGF of $Q = \boldsymbol{X}'\boldsymbol{X}$ when $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{I}_r)$. Construct an orthogonal $r \times r$ matrix $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{H}_2]$ whose first column equals $\boldsymbol{h}_1 = \boldsymbol{\mu}\left(\boldsymbol{\mu}'\boldsymbol{\mu}\right)^{-1/2}$. Note that $\boldsymbol{h}_1'\boldsymbol{\mu} = \lambda^{1/2}$ and $\boldsymbol{H}_2'\boldsymbol{\mu} = \boldsymbol{0}$. Define $\boldsymbol{Z} = \boldsymbol{H}'\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}^*, \boldsymbol{I}_q)$ where

$$\boldsymbol{\mu}^* = \boldsymbol{H}'\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{h}_1'\boldsymbol{\mu} \\ \boldsymbol{H}_2'\boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \lambda^{1/2} \\ \boldsymbol{0} \end{pmatrix} \begin{matrix} 1 \\ r-1 \end{matrix}.$$

It follows that $Q = \boldsymbol{X}'\boldsymbol{X} = \boldsymbol{Z}'\boldsymbol{Z} = Z_1^2 + \boldsymbol{Z}_2'\boldsymbol{Z}_2$ where $Z_1 \sim \mathrm{N}\left(\lambda^{1/2}, 1\right)$ and $\boldsymbol{Z}_2 \sim \mathrm{N}\left(0, \boldsymbol{I}_{r-1}\right)$ are

independent. Notice that $\boldsymbol{Z}_2'\boldsymbol{Z}_2 \sim \chi^2_{r-1}$ so has MGF $(1-2t)^{-(r-1)/2}$ by (5.31). The MGF of Z_1^2 is

$$\begin{aligned}
\mathbb{E}\left(\exp\left(tZ_1^2\right)\right) &= \int_{-\infty}^{\infty} \exp\left(tx^2\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(x-\sqrt{\lambda}\right)^2\right) dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(x^2(1-2t) - 2x\sqrt{\lambda} + \lambda\right)\right) dx \\
&= (1-2t)^{-1/2} \exp\left(-\frac{\lambda}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(u^2 - 2u\sqrt{\frac{\lambda}{1-2t}}\right)\right) du \\
&= (1-2t)^{-1/2} \exp\left(-\frac{\lambda t}{1-2t}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(u - \sqrt{\frac{\lambda}{1-2t}}\right)^2\right) du \\
&= (1-2t)^{-1/2} \exp\left(-\frac{\lambda t}{1-2t}\right)
\end{aligned}$$

where the third equality uses the change of variables $u = (1-2t)^{1/2} x$. Thus the MGF of $Q = Z_1^2 + \boldsymbol{Z}_2'\boldsymbol{Z}_2$ is

$$\begin{aligned}
\mathbb{E}\left(\exp\left(tQ\right)\right) &= \mathbb{E}\left(\exp\left(t\left(Z_1^2 + \boldsymbol{Z}_2'\boldsymbol{Z}_2\right)\right)\right) \\
&= \mathbb{E}\left(\exp\left(tZ_1^2\right)\right) \mathbb{E}\left(\exp\left(t\boldsymbol{Z}_2'\boldsymbol{Z}_2\right)\right) \\
&= (1-2t)^{-r/2} \exp\left(-\frac{\lambda t}{1-2t}\right). \qquad (5.32)
\end{aligned}$$

Second, we calculate the MGF of (5.4). It equals

$$\begin{aligned}
&\int_0^{\infty} \exp\left(tx\right) \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i f_{r+2i}(x) dx \\
&= \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i \int_0^{\infty} \exp\left(tx\right) f_{r+2i}(x) dx \\
&= \sum_{i=0}^{\infty} \frac{e^{-\lambda/2}}{i!} \left(\frac{\lambda}{2}\right)^i (1-2t)^{-(r+2i)/2} \\
&= e^{-\lambda/2} (1-2t)^{-r/2} \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{\lambda}{2(1-2t)}\right)^i \\
&= e^{-\lambda/2} (1-2t)^{-r/2} \exp\left(\frac{\lambda}{2(1-2t)}\right) \\
&= (1-2t)^{-r/2} \exp\left(\frac{\lambda t}{1-2t}\right) \qquad (5.33)
\end{aligned}$$

where the second equality uses (5.31), and the fourth uses $\exp(x) = \sum_{i=0}^{\infty} \frac{a^i}{i!}$. We can see that (5.32) equals (5.33), verifying that (5.4) is the density of Q as stated. ■

Proof of Theorem 5.3.3. The fact that $\boldsymbol{A} > 0$ means that we can write $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}'$ where $\boldsymbol{C}$ is non-singular (see Section A.9). Then $\boldsymbol{A}^{-1} = \boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}$ and by Theorem 5.2.3

$$\boldsymbol{C}^{-1}\boldsymbol{X} \sim \mathrm{N}\left(\boldsymbol{C}^{-1}\boldsymbol{\mu}, \boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{C}^{-1\prime}\right) = \mathrm{N}\left(\boldsymbol{C}^{-1}\boldsymbol{\mu}, \boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{C}'\boldsymbol{C}^{-1\prime}\right) = \mathrm{N}\left(\boldsymbol{\mu}^*, \boldsymbol{I}_r\right)$$

where $\boldsymbol{\mu}^* = \boldsymbol{C}^{-1}\boldsymbol{\mu}$. Thus by the definition of the non-central chi-square

$$\boldsymbol{X}'\boldsymbol{A}^{-1}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{X} = \left(\boldsymbol{C}^{-1}\boldsymbol{X}\right)'\left(\boldsymbol{C}^{-1}\boldsymbol{X}\right) \sim \chi^2_r\left(\boldsymbol{\mu}^{*\prime}\boldsymbol{\mu}^*\right).$$

Since

$$\boldsymbol{\mu}^{*\prime}\boldsymbol{\mu}^{*} = \boldsymbol{\mu}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{\mu} = \boldsymbol{\mu}'\boldsymbol{A}^{-1}\boldsymbol{\mu} = \lambda,$$

this equals $\chi^2_r(\lambda)$ as claimed. ■

Proof of Theorem 5.4.1. Using the simple law of iterated expectations, T has density

$$\begin{aligned}
f(x) &= \frac{d}{dx}\Pr\left(\frac{Z}{\sqrt{Q/r}} \le x\right) \\
&= \frac{d}{dx}\mathbb{E}\left\{Z \le x\sqrt{\frac{Q}{r}}\right\} \\
&= \frac{d}{dx}\mathbb{E}\left[\Pr\left(Z \le x\sqrt{\frac{Q}{r}} \mid Q\right)\right] \\
&= \mathbb{E}\frac{d}{dx}\Phi\left(x\sqrt{\frac{Q}{r}}\right) \\
&= \mathbb{E}\left(\phi\left(x\sqrt{\frac{Q}{r}}\right)\sqrt{\frac{Q}{r}}\right) \\
&= \int_0^\infty \left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{qx^2}{2r}\right)\right)\sqrt{\frac{q}{r}}\left(\frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}}q^{r/2-1}\exp\left(-q/2\right)\right)dq \\
&= \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}
\end{aligned}$$

using the gamma integral (5.28). ■

Proof of Theorem 5.4.2. Notice that for large r, by the properties of the logarithm

$$\log\left(\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}\right) = -\left(\frac{r+1}{2}\right)\log\left(1+\frac{x^2}{r}\right) \simeq -\left(\frac{r+1}{2}\right)\frac{x^2}{r} \to -\frac{x^2}{2},$$

the limit as $r \to \infty$, and thus

$$\lim_{r\to\infty}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)} = \exp\left(-\frac{x^2}{2}\right). \tag{5.34}$$

Using a property of the gamma function (5.29)

$$\lim_{n\to\infty}\frac{\Gamma(n+\alpha)}{\Gamma(n)n^\alpha} = 1$$

with $n = r/2$ and $\alpha = 1/2$ we find

$$\lim_{r\to\infty}\frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)} = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) = \phi(x).$$

■

Proof of Theorem 5.5.1. Let $U \sim \chi^2_m$ and $V \sim \chi^2_r$ be independent and set $S = U/V$. Let $f_m(u)$ be the χ^2_m density. By a similar argument as in the proof of Theorem 5.4.1, S has the density

function

$$\begin{aligned}
f_S(s) &= \mathbb{E}\left(f_m\left(sV\right)V\right) \\
&= \int_0^\infty f_m(sv) v f_r(v) dv \\
&= \frac{1}{2^{(m+r)/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{r}{2}\right)} \int_0^\infty (sv)^{m/2-1} e^{-sv/2} v^{r/2} e^{-v/2} dv \\
&= \frac{s^{m/2-1}}{2^{(m+r)/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{r}{2}\right)} \int_0^\infty v^{(m+r)/2-1} e^{-(s+1)v/2} dv \\
&= \frac{s^{m/2-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{r}{2}\right)(1+s)^{(m+r)/2}} \int_0^\infty t^{(m+r)/2-1} e^{-t} dt \\
&= \frac{s^{m/2-1}\Gamma\left(\frac{m+r}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{r}{2}\right)(1+s)^{(m+r)/2}}.
\end{aligned}$$

The fifth equality make the change-of variables $v = 2t/(1+s)$, and the sixth uses the definition of the Gamma function $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. Making the change-of-variables $x = sr/m$, we obtain the density as stated. ■

Proof of Theorem 5.5.2. The density of mF is

$$\frac{x^{m/2-1}\Gamma\left(\frac{m+r}{2}\right)}{r^{m/2}\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{r}{2}\right)\left(1+\frac{x}{r}\right)^{(m+r)/2}} \tag{5.35}$$

Using (5.29) with $n = r/2$ and $\alpha = m/2$ we have

$$\lim_{r\to\infty} \frac{\Gamma\left(\frac{m+r}{2}\right)}{r^{m/2}\Gamma\left(\frac{r}{2}\right)} = 2^{-m/2}$$

and similarly to (5.34) we have

$$\lim_{r\to\infty} \left(1+\frac{x}{r}\right)^{-\left(\frac{m+r}{2}\right)} = \exp\left(-\frac{x}{2}\right).$$

Together, (5.35) tends to

$$\frac{x^{m/2-1}\exp\left(-\frac{x}{2}\right)}{2^{m/2}\Gamma\left(\frac{m}{2}\right)}$$

which is the χ^2_m density. ■

Proof of Theorem 5.16.1. Since $\log(u)$ is concave we apply Jensen's inequality (B.5), take expectations are with respect to the true density $f\left(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}_0\right)$, and note that the density $f\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)$, integrates to 1 for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, to find that

$$\begin{aligned}
\mathbb{E}\left(\log \frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)} \mid \boldsymbol{X}\right) &\leq \log \mathbb{E}\left(\frac{L(\boldsymbol{\theta})}{L(\boldsymbol{\theta}_0)} \mid \boldsymbol{X}\right) \\
&= \log \int \cdots \int \left(\frac{\prod_{i=1}^n f\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right)}{\prod_{i=1}^n f\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0\right)}\right) \prod_{i=1}^n f\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0\right) d\boldsymbol{y}_1 \cdots d\boldsymbol{y}_n \\
&= \log \int \cdots \int \prod_{i=1}^n f\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}\right) d\boldsymbol{y}_1 \cdots d\boldsymbol{y}_n \\
&= \log 1 \\
&= 0.
\end{aligned}$$

This implies for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbb{E}(\log L(\boldsymbol{\theta})) \leq \mathbb{E}(\log L(\boldsymbol{\theta}_0))$. Hence $\boldsymbol{\theta}_0$ maximizes $\mathbb{E}(\log L(\boldsymbol{\theta}))$ as claimed. ■

Proof of Theorem 5.16.2. For part 1, Since the support of $\boldsymbol{y}$ does not depend on $\boldsymbol{\theta}$ we can exchange integration and differentiation:

$$\mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \mid \boldsymbol{X}\right) = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}\left(\log L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \mid \boldsymbol{X}\right).$$

Theorem 5.16.1 showed that $\mathbb{E}(\log L(\boldsymbol{\theta}))$ is maximized at $\boldsymbol{\theta}_0$, which has the first-order condition

$$\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E}\left(\log L(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \mid \boldsymbol{X}\right) = 0$$

as needed.

For part 2, using part 1 and the fact the observations are independent

$$\begin{aligned}
\boldsymbol{\mathcal{I}} &= \operatorname{var}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0) \mid \boldsymbol{X}\right) \\
&= \mathbb{E}\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0)\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0)\right)' \mid \boldsymbol{X}\right) \\
&= \sum_{i=1}^{n} \mathbb{E}\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)\right)' \mid \boldsymbol{x}_i\right)
\end{aligned}$$

which is the first equality.

For the second, observe that

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) = \frac{\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})}{f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})}$$

and

$$\begin{aligned}
\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) &= \frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})}{f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})} - \frac{\frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})'}{f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})^2} \\
&= \frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})}{f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})} - \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})'.
\end{aligned}$$

It follows that

$$\begin{aligned}
\boldsymbol{\mathcal{I}} &= \sum_{i=1}^{n} \mathbb{E}\left(\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)\right)' \mid \boldsymbol{x}_i\right) \\
&= -\sum_{i=1}^{n} \mathbb{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0) \mid \boldsymbol{x}_i\right) + \sum_{i=1}^{n} \mathbb{E}\left(\frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)}{f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)} \mid \boldsymbol{x}_i\right).
\end{aligned}$$

However, by exchanging integration and differentiation we can check that the second term is zero:

$$\begin{aligned}
\mathbb{E}\left(\frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)}{f(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)} \mid \boldsymbol{x}_i\right) &= \int \left(\frac{\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\boldsymbol{y} \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}}{f(\boldsymbol{y} \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)}\right) f(\boldsymbol{y}|\boldsymbol{\theta}_0)\, d\boldsymbol{y} \\
&= \int \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} f(\boldsymbol{y} \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} d\boldsymbol{y} \\
&= \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \int f(\boldsymbol{y} \mid \boldsymbol{x}_i, \boldsymbol{\theta}_0)\, d\boldsymbol{y}|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} 1 \\
&= 0
\end{aligned}$$

This establishes the second inequality. ■

Proof of Theorem 5.16.3 Let $\boldsymbol{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$ be the sample, let $f(\boldsymbol{Y}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i, \boldsymbol{\theta})$ denote the joint density of the sample, and note $\log L(\theta) = \log f(\boldsymbol{Y}, \boldsymbol{\theta})$. Set

$$\boldsymbol{S} = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}_0)$$

which by Theorem (5.16.2) has mean zero and variance $\boldsymbol{\mathcal{I}}$ conditional on $\boldsymbol{X}$. Write the estimator $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(\boldsymbol{Y})$ as a function of the data. Since $\widetilde{\boldsymbol{\theta}}$ is unbiased, for any $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} = \mathbb{E}\left(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{X}\right) = \int \widetilde{\boldsymbol{\theta}}(\boldsymbol{Y}) f(\boldsymbol{Y}, \boldsymbol{\theta}) \, d\boldsymbol{Y}.$$

Differentiating with respect to $\boldsymbol{\theta}$

$$\begin{aligned} \boldsymbol{I}_n &= \int \widetilde{\boldsymbol{\theta}}(\boldsymbol{Y}) \frac{\partial}{\partial \boldsymbol{\theta}'} f(\boldsymbol{Y}, \boldsymbol{\theta}) \, d\boldsymbol{Y} \\ &= \int \widetilde{\boldsymbol{\theta}}(\boldsymbol{Y}) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f(\boldsymbol{Y}, \boldsymbol{\theta}) \, f(\boldsymbol{Y}, \boldsymbol{\theta}) \, d\boldsymbol{Y}. \end{aligned}$$

Evaluating at $\boldsymbol{\theta}_0$ yields

$$\boldsymbol{I}_n = \mathbb{E}\left(\widetilde{\boldsymbol{\theta}} \boldsymbol{S}' \mid \boldsymbol{X}\right) = \mathbb{E}\left(\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \boldsymbol{S}' \mid \boldsymbol{X}\right) \tag{5.36}$$

the second equality since $\mathbb{E}(\boldsymbol{S} \mid \boldsymbol{X}) = 0$.

By the matrix Cauchy-Schwarz inequality (B.11), (5.36),and $\operatorname{var}(\boldsymbol{S} \mid \boldsymbol{X}) = \mathbb{E}(\boldsymbol{S}\boldsymbol{S}' \mid \boldsymbol{X}) = \boldsymbol{\mathcal{I}}$,

$$\begin{aligned} \operatorname{var}\left(\widetilde{\boldsymbol{\theta}} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \mid \boldsymbol{X}\right) \\ &\geq \mathbb{E}\left(\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \boldsymbol{S}' \mid \boldsymbol{X}\right) \left(\mathbb{E}(\boldsymbol{S}\boldsymbol{S}' \mid \boldsymbol{X})\right)^{-1} \mathbb{E}\left(\boldsymbol{S}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \mid \boldsymbol{X}\right) \\ &= \left(\mathbb{E}(\boldsymbol{S}\boldsymbol{S}' \mid \boldsymbol{X})\right)^{-1} \\ &= \boldsymbol{\mathcal{I}}^{-1} \end{aligned}$$

as stated. ■

Exercises

Exercise 5.1 For the standard normal density $\phi(x)$, show that $\phi'(x) = -x\phi(x)$.

Exercise 5.2 Use the result in Exercise 5.1 and integration by parts to show that for $X \sim \mathrm{N}(0,1)$, $\mathbb{E}X^2 = 1$.

Exercise 5.3 Use the results in Exercises 5.1 and 5.2, plus integration by parts, to show that for $X \sim \mathrm{N}(0,1)$, $\mathbb{E}X^4 = 3$.

Exercise 5.4 Show that the moment generating function (mgf) of $X \sim \mathrm{N}(0,1)$ is $m(t) = \mathbb{E}(\exp(tX)) = \exp\left(t^2/2\right)$. (For the definition of the mgf see Section 2.31).

Exercise 5.5 Use the mgf from Exercise 5.4 to verify that for $X \sim \mathrm{N}(0,1)$, $\mathbb{E}\left(X^2\right) = m''(0) = 1$ and $\mathbb{E}\left(X^4\right) = m^{(4)}(0) = 3$.

Exercise 5.6 Write the multivariate $\mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ density as the product of $\mathrm{N}(0,1)$ density functions. That is, show that

$$\frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{\boldsymbol{x}'\boldsymbol{x}}{2}\right) = \phi(x_1)\cdots\phi(x_k).$$

Exercise 5.7 Show that the mgf of $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ is $\mathbb{E}(\exp(\boldsymbol{t}'\boldsymbol{X})) = \exp\left(\frac{1}{2}\boldsymbol{t}'\boldsymbol{t}\right)$.

Hint: Use Exercise 5.4 and the fact that the elements of $\boldsymbol{X}$ are independent.

Exercise 5.8 Show that the mgf of $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$M(\boldsymbol{t}) = \mathbb{E}\left(\exp\left(\boldsymbol{t}'\boldsymbol{X}\right)\right) = \exp\left(\boldsymbol{t}'\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

Hint: Write $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$ where $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_k)$.

Exercise 5.9 Show that the characteristic function of $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$C(\boldsymbol{t}) = \mathbb{E}\left(\exp\left(\mathrm{i}\boldsymbol{t}'\boldsymbol{X}\right)\right) = \exp\left(\mathrm{i}\boldsymbol{\mu}'\boldsymbol{\lambda} - \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

For the definition of the characteristic function see Section 2.31

Hint: For $X \sim \mathrm{N}(0,1)$, establish $\mathbb{E}(\exp(\mathrm{i}tX)) = \exp\left(-\frac{1}{2}t^2\right)$ by integration. Then generalize to $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using the same steps as in Exercises 5.7 and 5.8.

Exercise 5.10 Show that if $Q \sim \chi^2_r$, then $\mathbb{E}(Q) = r$ and $\mathrm{var}(Q) = 2r$.

Hint: Use the representation $Q = \sum_{i=1}^n X_i^2$ with X_i independent $\mathrm{N}(0,1)$.

Exercise 5.11 Show that if $Q \sim \chi^2_k(\lambda)$, then $\mathbb{E}(Q) = k + \lambda$.

Exercise 5.12 Suppose X_i are independent $\mathrm{N}\left(\mu_i, \sigma_i^2\right)$. Find the distribution of the weighted sum $\sum_{i=1}^n w_i X_i$.

Exercise 5.13 Show that if $\boldsymbol{e} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_n\sigma^2\right)$ and $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$ then $\boldsymbol{u} = \boldsymbol{H}'\boldsymbol{e} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_n\sigma^2\right)$.

Exercise 5.14 Show that if $\boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}'$ then $\boldsymbol{u} = \boldsymbol{A}^{-1}\boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_n)$.

Exercise 5.15 Show that $\widehat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \log L(\boldsymbol{\theta}) = \mathrm{argmax}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})$.

Exercise 5.16 For the regression in-sample predicted values $\widehat{y}_i$ show that $\widehat{y}_i|_{\boldsymbol{X}} \sim \mathrm{N}\left(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2 h_{ii}\right)$ where h_{ii} are the leverage values (3.25).

Exercise 5.17 In the normal regression model, show that the leave-one out prediction errors $\widetilde{e}_i$ and the standardized residuals $\bar{e}_i$ are independent of $\widehat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$.

Hint: Use (3.46) and (4.26).

Exercise 5.18 In the normal regression model, show that the robust covariance matrices $\widehat{\boldsymbol{V}}^W_{\widehat{\boldsymbol{\beta}}}$, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, and $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ are independent of the OLS estimate $\widehat{\boldsymbol{\beta}}$, conditional on $\boldsymbol{X}$.

Exercise 5.19 Let $F(u)$ be the distribution function of a random variable X whose density is symmetric about zero. (This includes the standard normal and the student t.) Show that $F(-u) = 1 - F(u)$.

Exercise 5.20 Let $C_\beta = [L, U]$ be a $1-\alpha$ confidence interval for β, and consider the transformation $\theta = g(\beta)$ where $g(\cdot)$ is monotonically increasing. Consider the confidence interval $C_\theta = [g(L), g(U)]$ for θ. Show that $\Pr\left(\theta \in C_\theta\right) = \Pr\left(\beta \in C_\beta\right)$. Use this result to develop a confidence interval for σ.

Exercise 5.21 Show that the test "Reject $\mathbb{H}_0$ if $LR \geq c_1$" for LR defined in (5.22), and the test "Reject $\mathbb{H}_0$ if $F \geq c_2$" for F defined in (5.23), yield the same decisions if $c_2 = \left(\exp(c_1/n) - 1\right)(n - k)/q$. Why does this mean that the two tests are *equivalent*?

Exercise 5.22 Show (5.24).

Exercise 5.23 In the normal regression model, let s^2 be the unbiased estimator of the error variance σ^2 from (4.30).

(a) Show that $\operatorname{var}\left(s^2\right) = 2\sigma^4/(n-k)$.

(b) Show that $\operatorname{var}\left(s^2\right)$ is strictly larger than the Cramér-Rao Lower Bound for σ^2.

Chapter 6

An Introduction to Large Sample Asymptotics

6.1 Introduction

For inference (confidence intervals and hypothesis testing) on unknown parameters we need sampling distributions, either exact or approximate, of estimates and other statistics.

In Chapter 4 we derived the mean and variance of the least-squares estimator in the context of the linear regression model, but this is not a complete description of the sampling distribution and is thus not sufficient for inference. Furthermore, the theory does not apply in the context of the linear projection model, which is more relevant for empirical applications.

In Chapter 5 we derived the exact sampling distribution of the OLS estimator, t-statistics, and F-statistics for the normal regression model, allowing for inference. But these results are narrowly confined to the normal regression model, which requires the unrealistic assumption that the regression error is normally distributed and independent of the regressors. Perhaps we can view these results as some sort of approximation to the sampling distributions without requiring the assumption of normality, but how can we be precise about this?

To illustrate the situation with an example, let y_i and x_i be drawn from the joint density

$$f(x,y) = \frac{1}{2\pi xy}\exp\left(-\frac{1}{2}\left(\log y - \log x\right)^2\right)\exp\left(-\frac{1}{2}\left(\log x\right)^2\right)$$

and let $\widehat{\beta}$ be the slope coefficient estimate from a least-squares regression of y_i on x_i and a constant. Using simulation methods, the density function of $\widehat{\beta}$ was computed and plotted in Figure 6.1 for sample sizes of $n = 25$, $n = 100$ and $n = 800$. The vertical line marks the true projection coefficient.

From the figure we can see that the density functions are dispersed and highly non-normal. As the sample size increases the density becomes more concentrated about the population coefficient. Is there a simple way to characterize the sampling distribution of $\widehat{\beta}$?

In principle the sampling distribution of $\widehat{\beta}$ is a function of the joint distribution of (y_i, x_i) and the sample size n, but in practice this function is extremely complicated so it is not feasible to analytically calculate the exact distribution of $\widehat{\beta}$ except in very special cases. Therefore we typically rely on approximation methods.

In this chapter we introduce asymptotic theory, which approximates by taking the limit of the finite sample distribution as the sample size n tends to infinity. It is important to understand that this is an approximation technique, as the asymptotic distributions are used to assess the finite sample distributions of our estimators in actual practical samples. The primary tools of asymptotic theory are the weak law of large numbers (WLLN), central limit theorem (CLT), and continuous mapping theorem (CMT). With these tools we can approximate the sampling distributions of most econometric estimators.

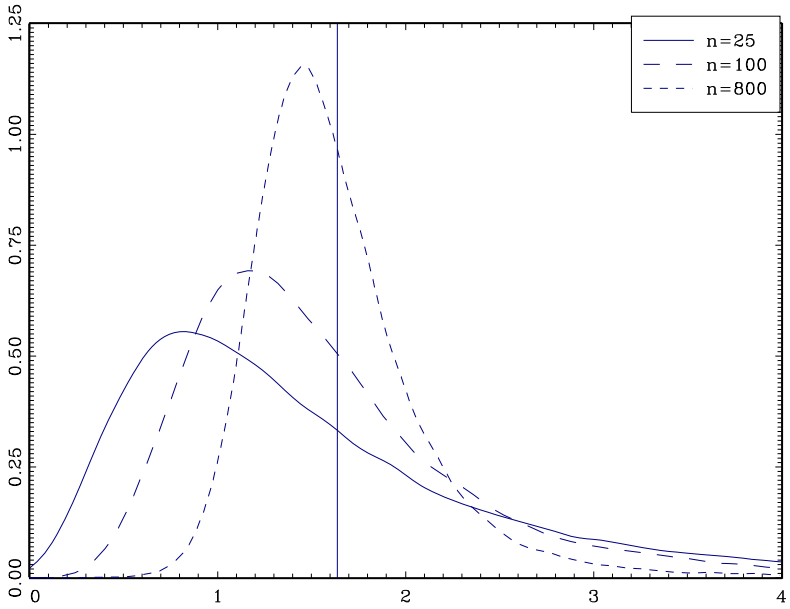


Figure 6.1: Sampling Density of $\hat{\beta}$

In this chapter we provide a concise summary. It will be useful for most students to review this material, even if most is familiar.

6.2 Asymptotic Limits

"Asymptotic analysis" is a method of approximation obtained by taking a suitable limit. There is more than one method to take limits, but the most common is to take the limit of the sequence of sampling distributions as the sample size tends to positive infinity, written "as $n \to \infty$." It is not meant to be interpreted literally, but rather as an approximating device.

The first building block for asymptotic analysis is the concept of a limit of a sequence.

Definition 6.2.1 *A sequence a_n has the* ***limit*** *a, written $a_n \longrightarrow a$ as $n \to \infty$, or alternatively as $\lim_{n\to\infty} a_n = a$, if for all $\delta > 0$ there is some $n_\delta < \infty$ such that for all $n \geq n_\delta$, $|a_n - a| \leq \delta$.*

In words, a_n has the limit a if the sequence gets closer and closer to a as n gets larger. If a sequence has a limit, that limit is unique (a sequence cannot have two distinct limits). If a_n has the limit a, we also say that a_n **converges** to a as $n \to \infty$.

Not all sequences have limits. For example, the sequence $\{1, 2, 1, 2, 1, 2, ...\}$ does not have a limit. It is therefore sometimes useful to have a more general definition of limits which always exist, and these are the limit superior and limit inferior of a sequence.

Definition 6.2.2 $\liminf_{n\to\infty} a_n \stackrel{def}{=} \lim_{n\to\infty} \inf_{m\geq n} a_m$

Definition 6.2.3 $\limsup_{n\to\infty} a_n \stackrel{def}{=} \lim_{n\to\infty} \sup_{m\geq n} a_m$

The limit inferior and limit superior always exist (including $\pm\infty$ as possibilities), and equal when the limit exists. In the example given earlier, the limit inferior of $\{1, 2, 1, 2, 1, 2, ...\}$ is 1, and the limit superior is 2.

6.3 Convergence in Probability

A sequence of numbers may converge to a limit, but what about a sequence of random variables? For example, consider a sample mean $\overline{y} = n^{-1} \sum_{i=1}^{n} y_i$ based on an random sample of n observations. As n increases, the distribution of $\overline{y}$ changes. In what sense can we describe the "limit" of $\overline{y}$? In what sense does it converge?

Since $\overline{y}$ is a random variable, we cannot directly apply the deterministic concept of a sequence of numbers. Instead, we require a definition of convergence which is appropriate for random variables. There are more than one such definition, but the most commonly used is called convergence in probability.

Definition 6.3.1 *A random variable $z_n \in \mathbb{R}$* ***converges in probability*** *to z as $n \to \infty$, denoted $z_n \xrightarrow{p} z$, or alternatively $\operatorname{plim}_{n\to\infty} z_n = z$, if for all $\delta > 0$,*

$$\lim_{n\to\infty} \Pr\left(|z_n - z| \leq \delta\right) = 1. \tag{6.1}$$

We call z the ***probability limit*** *(or* ***plim****) of z_n.*

The definition looks quite abstract, but it formalizes the concept of a sequence of random variables concentrating about a point. The event $\{|z_n - z| \leq \delta\}$ occurs when z_n is within δ of the point z. $\Pr\left(|z_n - z| \leq \delta\right)$ is the probability of this event – that z_n is within δ of the point z. Equation (6.1) states that this probability approaches 1 as the sample size n increases. The definition of convergence in probability requires that this holds for any δ. So for any small interval about z the distribution of z_n concentrates within this interval for large n.

You may notice that the definition concerns the distribution of the random variables z_n, not their realizations. Furthermore, notice that the definition uses the concept of a conventional (deterministic) limit, but the latter is applied to a sequence of probabilities, not directly to the random variables z_n or their realizations.

Two comments about the notation are worth mentioning. First, it is conventional to write the convergence symbol as $\xrightarrow{p}$ where the "p" above the arrow indicates that the convergence is "in probability". You should try and adhere to this notation, and not simply write $z_n \longrightarrow z$. Second, it is important to include the phrase "as $n \to \infty$" to be specific about how the limit is obtained.

A common mistake is to confuse convergence in probability with convergence in expectation:

$$\mathbb{E}\left(z_n\right) \longrightarrow \mathbb{E}\left(z\right). \tag{6.2}$$

They are related but distinct concepts. Neither (6.1) nor (6.2) implies the other.

To see the distinction it might be helpful to think through a stylized example. Consider a discrete random variable z_n which takes the value 0 with probability $1 - n^{-1}$ and the value $a_n \neq 0$ with probability n^{-1}, or

$$\begin{aligned} \Pr\left(z_n = 0\right) &= 1 - \frac{1}{n} \\ \Pr\left(z_n = a_n\right) &= \frac{1}{n}. \end{aligned} \tag{6.3}$$

In this example the probability distribution of z_n concentrates at zero as n increases, regardless of the sequence a_n. You can check that $z_n \xrightarrow{p} 0$ as $n \to \infty$.

In this example we can also calculate that the expectation of z_n is

$$\mathbb{E}(z_n) = \frac{a_n}{n}.$$

Despite the fact that z_n converges in probability to zero, its expectation will not decrease to zero unless $a_n/n \to 0$. If a_n diverges to infinity at a rate equal to n (or faster) then $\mathbb{E}(z_n)$ will not converge to zero. For example, if $a_n = n$, then $\mathbb{E}(z_n) = 1$ for all n, even though $z_n \xrightarrow{p} 0$. This example might seem a bit artificial, but the point is that the concepts of convergence in probability and convergence in expectation are distinct, so it is important not to confuse one with the other.

Another common source of confusion with the notation surrounding probability limits is that the expression to the right of the arrow "$\xrightarrow{p}$" must be free of dependence on the sample size n. Thus expressions of the form "$z_n \xrightarrow{p} c_n$" are notationally meaningless and should not be used.

6.4 Weak Law of Large Numbers

In large samples we expect parameter estimates to be close to the population values. For example, in Section 4.3 we saw that the sample mean $\overline{y}$ is unbiased for $\mu = \mathbb{E}(y)$ and has variance σ^2/n. As n gets large its variance decreases and thus the distribution of $\overline{y}$ concentrates about the population mean μ. It turns out that this implies that the sample mean converges in probability to the population mean.

When y has a finite variance there is a fairly straightforward proof by applying Chebyshev's inequality.

Theorem 6.4.1 ***Chebyshev's Inequality.*** *For any random variable z_n and constant $\delta > 0$*

$$\Pr(|z_n - \mathbb{E}z_n| \geq \delta) \leq \frac{\operatorname{var}(z_n)}{\delta^2}.$$

Chebyshev's inequality is terrifically important in asymptotic theory. While its proof is a technical exercise in probability theory, it is quite simple so we discuss it forthwith. Let $F_n(u)$ denote the distribution of $z_n - \mathbb{E}z_n$. Then

$$\Pr(|z_n - \mathbb{E}z_n| \geq \delta) = \Pr\left((z_n - \mathbb{E}z_n)^2 \geq \delta^2\right) = \int_{\{u^2 \geq \delta^2\}} dF_n(u).$$

The integral is over the event $\left\{u^2 \geq \delta^2\right\}$, so that the inequality $1 \leq \dfrac{u^2}{\delta^2}$ holds throughout. Thus

$$\int_{\{u^2 \geq \delta^2\}} dF_n(u) \leq \int_{\{u^2 \geq \delta^2\}} \frac{u^2}{\delta^2} dF_n(u) \leq \int \frac{u^2}{\delta^2} dF_n(u) = \frac{\mathbb{E}(z_n - \mathbb{E}z_n)^2}{\delta^2} = \frac{\operatorname{var}(z_n)}{\delta^2},$$

which establishes the desired inequality.

Applied to the sample mean $\overline{y}$ which has variance σ^2/n, Chebyshev's inequality shows that for any $\delta > 0$

$$\Pr(|\overline{y} - \mathbb{E}(\overline{y})| \geq \delta) \leq \frac{\sigma^2/n}{\delta^2}.$$

For fixed σ^2 and δ, the bound on the right-hand-side shrinks to zero as $n \to \infty$. (Specifically, for any $\varepsilon > 0$ set $n \geq \sigma^2 / \left(\delta^2 \varepsilon\right)$. Then the right-hand-side is less than or equal to ε.) Thus the probability that $\overline{y}$ is within δ of $\mathbb{E}(\overline{y}) = \mu$ approaches 1 as n gets large, or

$$\lim_{n \to \infty} \Pr\left(|\overline{y} - \mu| < \delta\right) = 1.$$

This means that $\overline{y}$ converges in probability to μ as $n \to \infty$.

This result is called the **weak law of large numbers**. Our derivation assumed that y has a finite variance, but with a more careful derivation all that is necessary is a finite mean.

Theorem 6.4.2 ***Weak Law of Large Numbers (WLLN)***
If y_i are independent and identically distributed and $\mathbb{E}|y| < \infty$, then as $n \to \infty$,

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \xrightarrow{p} \mathbb{E}(y).$$

The proof of Theorem 6.4.2 is presented in Section 6.16.

The WLLN shows that the estimator $\overline{y}$ converges in probability to the true population mean μ. In general, an estimator which converges in probability to the population value is called **consistent**.

Definition 6.4.1 *An estimator $\widehat{\theta}$ of a parameter θ is **consistent** if $\widehat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.*

Theorem 6.4.3 *If y_i are independent and identically distributed and $\mathbb{E}|y| < \infty$, then $\widehat{\mu} = \overline{y}$ is consistent for the population mean μ.*

Consistency is a good property for an estimator to possess. It means that for any given data distribution, there is a sample size n sufficiently large such that the estimator $\widehat{\theta}$ will be arbitrarily close to the true value θ with high probability. The theorem does not tell us, however, how large this n has to be. Thus the theorem does not give practical guidance for empirical practice. Still, it is a minimal property for an estimator to be considered a "good" estimator, and provides a foundation for more useful approximations.

6.5 Almost Sure Convergence and the Strong Law*

Convergence in probability is sometimes called **weak convergence**. A related concept is **almost sure convergence**, also known as **strong convergence**. (In probability theory the term "almost sure" means "with probability equal to one". An event which is random but occurs with probability equal to one is said to be **almost sure**.)

Definition 6.5.1 *A random variable $z_n \in \mathbb{R}$* ***converges almost surely*** *to z as $n \to \infty$, denoted $z_n \xrightarrow{a.s.} z$, if for every $\delta > 0$*

$$\Pr\left(\lim_{n\to\infty} |z_n - z| \leq \delta\right) = 1. \tag{6.4}$$

The convergence (6.4) is stronger than (6.1) because it computes the probability of a limit rather than the limit of a probability. Almost sure convergence is stronger than convergence in probability in the sense that $z_n \xrightarrow{a.s.} z$ implies $z_n \xrightarrow{p} z$.

In the example (6.3) of Section 6.3, the sequence z_n converges in probability to zero for any sequence a_n, but this is not sufficient for z_n to converge almost surely. In order for z_n to converge to zero almost surely, it is necessary that $a_n \to 0$.

In the random sampling context the sample mean can be shown to converge almost surely to the population mean. This is called the **strong law of large numbers**.

Theorem 6.5.1 ***Strong Law of Large Numbers (SLLN)***
If y_i are independent and identically distributed and $\mathbb{E}|y| < \infty$, then as $n \to \infty$,

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \xrightarrow{a.s.} \mathbb{E}(y).$$

The proof of the SLLN is technically quite advanced so is not presented here. For a proof see Billingsley (1995, Theorem 22.1) or Ash (1972, Theorem 7.2.5).

The WLLN is sufficient for most purposes in econometrics, so we will not use the SLLN in this text.

6.6 Vector-Valued Moments

Our preceding discussion focused on the case where y is real-valued (a scalar), but nothing important changes if we generalize to the case where $\boldsymbol{y} \in \mathbb{R}^m$ is a vector. To fix notation, the elements of $\boldsymbol{y}$ are

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

The population mean of $\boldsymbol{y}$ is just the vector of marginal means

$$\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y}) = \begin{pmatrix} \mathbb{E}(y_1) \\ \mathbb{E}(y_2) \\ \vdots \\ \mathbb{E}(y_m) \end{pmatrix}.$$

When working with random vectors $\boldsymbol{y}$ it is convenient to measure their magnitude by their Euclidean length or Euclidean norm

$$\|\boldsymbol{y}\| = \left(y_1^2 + \cdots + y_m^2\right)^{1/2}.$$

In vector notation we have

$$\|\boldsymbol{y}\|^2 = \boldsymbol{y}'\boldsymbol{y}.$$

It turns out that it is equivalent to describe finiteness of moments in terms of the Euclidean norm of a vector or all individual components.

Theorem 6.6.1 *For $\boldsymbol{y} \in \mathbb{R}^m$, $\mathbb{E}\|\boldsymbol{y}\| < \infty$ if and only if $\mathbb{E}|y_j| < \infty$ for $j = 1, ..., m$.*

The $m \times m$ variance matrix of $\boldsymbol{y}$ is

$$\boldsymbol{V} = \operatorname{var}(\boldsymbol{y}) = \mathbb{E}\left((\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'\right).$$

$\boldsymbol{V}$ is often called a variance-covariance matrix. You can show that the elements of $\boldsymbol{V}$ are finite if $\mathbb{E}\|\boldsymbol{y}\|^2 < \infty$.

A random sample $\{\boldsymbol{y}_1, ..., \boldsymbol{y}_n\}$ consists of n observations of independent and identically distributed draws from the distribution of $\boldsymbol{y}$. (Each draw is an m-vector.) The vector sample mean

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i = \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \vdots \\ \overline{y}_m \end{pmatrix}$$

is the vector of sample means of the individual variables.

Convergence in probability of a vector can be defined as convergence in probability of all elements in the vector. Thus $\overline{\boldsymbol{y}} \xrightarrow{p} \boldsymbol{\mu}$ if and only if $\overline{y}_j \xrightarrow{p} \mu_j$ for $j = 1, ..., m$. Since the latter holds if $\mathbb{E}|y_j| < \infty$ for $j = 1, ..., m$, or equivalently $\mathbb{E}\|\boldsymbol{y}\| < \infty$, we can state this formally as follows.

Theorem 6.6.2 ***WLLN for random vectors***
If $\boldsymbol{y}_i$ are independent and identically distributed and $\mathbb{E}\|\boldsymbol{y}\| < \infty$, then as $n \to \infty$,

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \xrightarrow{p} \mathbb{E}(\boldsymbol{y}).$$

6.7 Convergence in Distribution

The WLLN is a useful first step, but does not give an approximation to the distribution of an estimator. A **large-sample** or **asymptotic** approximation can be obtained using the concept of **convergence in distribution**.

We say that a sequence of random vectors $\boldsymbol{z}_n$ converges in distribution if the sequence of distribution functions $F_n(\boldsymbol{u}) = \Pr(\boldsymbol{z}_n \leq \boldsymbol{u})$ converges to a limit distribution function.

Definition 6.7.1 *Let $\boldsymbol{z}_n$ be a random vector with distribution $F_n(\boldsymbol{u}) = \Pr(\boldsymbol{z}_n \leq \boldsymbol{u})$. We say that $\boldsymbol{z}_n$ **converges in distribution** to $\boldsymbol{z}$ as $n \to \infty$, denoted $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$, if for all $\boldsymbol{u}$ at which $F(\boldsymbol{u}) = \Pr(\boldsymbol{z} \leq \boldsymbol{u})$ is continuous, $F_n(\boldsymbol{u}) \to F(\boldsymbol{u})$ as $n \to \infty$.*

Under these conditions, it is also said that F_n **converges weakly** to F. It is common to refer to $\boldsymbol{z}$ and its distribution $F(u)$ as the **asymptotic distribution**, **large sample distribution**, or **limit distribution** of $\boldsymbol{z}_n$.

When the limit distribution $\boldsymbol{z}$ is degenerate (that is, $\Pr(\boldsymbol{z} = \boldsymbol{c}) = 1$ for some $\boldsymbol{c}$) we can write the convergence as $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{c}$, which is equivalent to convergence in probability, $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{c}$.

Technically, in most cases of interest it is difficult to establish the limit distributions of sample statistics $\boldsymbol{z}_n$ by working directly with their distribution function. It turns out that in most cases it is easier to work with their characteristic function $C_n(\boldsymbol{\lambda}) = \mathbb{E}\left(\exp\left(\mathrm{i}\boldsymbol{\lambda}'\boldsymbol{z}_n\right)\right)$, which is a transformation of the distribution. (See Section 2.31 for the definition.) While this is more technical than needed for most applied economists, we introduce this material to give a complete reference for large sample approximations.

The characteristic function $C_n(\boldsymbol{t})$ completely describes the distribution of $\boldsymbol{z}_n$. It therefore seems reasonable to expect that if $C_n(\boldsymbol{t})$ converges to a limit function $C(\boldsymbol{t})$, then the the distribution of $\boldsymbol{z}_n$ converges as well. This turns out to be true, and is known as Lévy's continuity theorem.

Theorem 6.7.1 ***Lévy's Continuity Theorem.*** *$\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ if and only if $\mathbb{E}(\exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{z}_n)) \to \mathbb{E}(\exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{z}))$ for every $\boldsymbol{t} \in \mathbb{R}^k$.*

While this result seems quite intuitive, a rigorous proof is quite advanced and so is not presented here. See Van der Vaart (2008) Theorem 2.13.

Finally, we mention a standard trick which is commonly used to establish multivariate convergence results.

Theorem 6.7.2 ***Cramér-Wold Device.*** *$\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ if and only if $\boldsymbol{\lambda}'\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{\lambda}'\boldsymbol{z}$ for every $\boldsymbol{\lambda} \in \mathbb{R}^k$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$.*

We present a proof in Section 6.16 which is a simple application of Lévy's continuity theorem.

6.8 Central Limit Theorem

We would like to obtain a distributional approximation to the sample mean $\overline{y}$. We start under the random sampling assumption so that the observations are independent and identically distributed, and have a finite mean $\mu = \mathbb{E}(y)$ and variance $\sigma^2 = \mathrm{var}(y)$.

Let's start by finding the asymptotic distribution of $\overline{y}$, in the sense that $\overline{y} \xrightarrow{d} z$ for some random variable z. From the WLLN we know that $\overline{y} \xrightarrow{p} \mu$. Since convergence in probability to a constant is the same as convergence in distribution, this means that $\overline{y} \xrightarrow{d} \mu$ as well. This is not a useful distributional result as the limit distribution is a constant. To obtain a non-degenerate distribution

we need to rescale $\overline{y}$. Recall that $\operatorname{var}(\overline{y}-\mu)=\sigma^2/n$, which means that $\operatorname{var}(\sqrt{n}(\overline{y}-\mu))=\sigma^2$. This suggests renormalizing the statistic as

$$z_n=\sqrt{n}(\overline{y}-\mu).$$

Notice that $\mathbb{E}(z_n)=0$ and $\operatorname{var}(z_n)=\sigma^2$. This shows that the mean and variance have been stabilized. We now seek to determine the asymptotic distribution of z_n.

The answer is provided by the central limit theorem (CLT) which states that standardized sample averages converge in distribution to normal random vectors. There are several versions of the CLT. The most basic is the case where the observations are independent and identically distributed.

Theorem 6.8.1 ***Lindeberg–Lévy Central Limit Theorem.*** *If y_i are independent and identically distributed and $\mathbb{E}\left(y_i^2\right)<\infty$, then as $n\to\infty$*

$$\sqrt{n}(\overline{y}-\mu)\xrightarrow{d}\mathrm{N}\left(0,\sigma^2\right)$$

where $\mu=\mathbb{E}(y)$ and $\sigma^2=\mathbb{E}(y_i-\mu)^2$.

The proof of the CLT is rather technical (so is presented in Section 6.16) but at the core is a quadratic approximation of the log of the characteristic function.

As we discussed above, in finite samples the standardized sum $z_n=\sqrt{n}(\overline{y}_n-\mu)$ has mean zero and variance σ^2. What the CLT adds is that z_n is also approximately normally distributed, and that the normal approximation improves as n increases.

The CLT is one of the most powerful and mysterious results in statistical theory. It shows that the simple process of averaging induces normality. The first version of the CLT (for the number of heads resulting from many tosses of a fair coin) was established by the French mathematician Abraham de Moivre in an article published in 1733. This was extended to cover an approximation to the binomial distribution in 1812 by Pierre-Simon Laplace in his book *Théorie Analytique des Probabilités*, and the most general statements are credited to articles by the Russian mathematician Aleksandr Lyapunov (1901) and the Finnish mathematician Jarl Waldemar Lindeberg (1920, 1922). The above statement is known as the classic (or Lindeberg-Lévy) CLT due to contributions by Lindeberg (1920) and the French mathematician Paul Pierre Lévy.

A more general version which allows heterogeneous distributions was provided by Lindeberg (1922). The following is the most general statement.

Theorem 6.8.2 ***Lindeberg-Feller Central Limit Theorem.*** *Suppose y_{ni} are independent but not necessarily identically distributed with finite means $\mu_{ni}=\mathbb{E}(y_{ni})$ and variances $\sigma_{ni}^2=\mathbb{E}(y_{ni}-\mu_{ni})^2$. Set $\overline{\sigma}_n^2=n^{-1}\sum_{i=1}^n\sigma_{ni}^2$. If $\overline{\sigma}_n^2>0$ and for all $\varepsilon>0$*

$$\lim_{n\to\infty}\frac{1}{n\overline{\sigma}_n^2}\sum_{i=1}^n\mathbb{E}\left(\left(y_{ni}-\mu_{ni}\right)^2\mathbb{1}\left(\left(y_{ni}-\mu_{ni}\right)^2\geq\varepsilon n\overline{\sigma}_n^2\right)\right)=0 \tag{6.5}$$

then as $n\to\infty$

$$\frac{\sqrt{n}(\overline{y}-\mathbb{E}(\overline{y}))}{\overline{\sigma}_n^{1/2}}\xrightarrow{d}\mathrm{N}(0,1).$$

The proof of the Lindeberg-Feller CLT is substantially more technical, so we do not present it here. See Billingsley (1995, Theorem 27.2).

The Lindeberg-Feller CLT is quite general as it puts minimal conditions on the sequence of means and variances. The key assumption is equation (6.5) which is known as **Lindeberg's Condition**. In its raw form it is difficult to interpret. The intuition for (6.5) is that it excludes any single observation from dominating the asymptotic distribution. Since (6.5) is quite abstract, in most contexts we use more elementary conditions which are simpler to interpret.

One such alternative is called **Lyapunov's condition**: For some $\delta > 0$

$$\lim_{n\to\infty} \frac{1}{n^{1+\delta/2}\overline{\sigma}_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left(|y_{ni} - \mu_{ni}|^{2+\delta}\right) = 0. \tag{6.6}$$

Lyapunov's condition implies Lindeberg's condition, and hence the CLT. Indeed, the left-side of (6.5) is bounded by

$$\begin{aligned}
&\lim_{n\to\infty} \frac{1}{n\overline{\sigma}_n^2} \sum_{i=1}^{n} \mathbb{E}\left(\frac{|y_{ni} - \mu_{ni}|^{2+\delta}}{|y_{ni} - \mu_{ni}|^{\delta}} 1\left(|y_{ni} - \mu_{ni}|^2 \geq \varepsilon n\overline{\sigma}_n^2\right)\right) \\
&\leq \lim_{n\to\infty} \frac{1}{\varepsilon^{\delta/2} n^{1+\delta/2}\overline{\sigma}_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left(|y_{ni} - \mu_{ni}|^{2+\delta}\right) \\
&= 0
\end{aligned}$$

by (6.6).

Lyapunov's condition is still awkward to interpret. A still simpler condition is a uniform moment bound: For some $\delta > 0$

$$\sup_{n,i} \mathbb{E}\,|y_{ni}|^{2+\delta} < \infty. \tag{6.7}$$

This is typically combined with the lower variance bound

$$\liminf_{n\to\infty} \overline{\sigma}_n^2 > 0. \tag{6.8}$$

These bounds together imply Lyapunov's condition. To see this, (6.7) and (6.8) imply there is some $C < \infty$ such that $\sup_{n,i} \mathbb{E}\,|y_{ni}|^{2+\delta} \leq C$ and $\liminf_{n\to\infty} \overline{\sigma}_n^2 \geq C^{-1}$. Without loss of generality assume $\mu_{ni} = 0$. Then the left side of (6.6) is bounded by

$$\lim_{n\to\infty} \frac{C^{2+\delta/2}}{n^{\delta/2}} = 0,$$

so Lyapunov's condition holds and hence the CLT.

An alternative to (6.8) is to assume that the average variance $\overline{\sigma}_n^2$ converges to a constant, that is,

$$\overline{\sigma}_n^2 = n^{-1} \sum_{i=1}^{n} \sigma_{ni}^2 \to \sigma^2 < \infty. \tag{6.9}$$

This assumption is reasonable in many applications.

We now state the simplest and most commonly used version of a heterogeneous CLT based on the Lindeberg-Feller Theorem.

Theorem 6.8.3 *Suppose y_{ni} are independent but not necessarily identically distributed. If (6.7) and (6.9) hold, then as $n \to \infty$*

$$\sqrt{n}\left(\overline{y} - \mathbb{E}\left(\overline{y}\right)\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\right). \tag{6.10}$$

One advantage of Theorem 6.8.3 is that it allows $\sigma^2 = 0$ (unlike Theorem 6.8.2).

6.9 Multivariate Central Limit Theorem

Multivariate central limit theory applies when we consider vector-valued observations $\boldsymbol{y}_i$ and sample averages $\overline{\boldsymbol{y}}$. In the i.i.d. case we know that the mean of $\overline{\boldsymbol{y}}$ is the mean vector $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y})$ and its variance is $n^{-1}\boldsymbol{V}$ where $\boldsymbol{V} = \mathbb{E}\left((\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'\right)$. Again we wish to transform $\overline{\boldsymbol{y}}$ so that its mean and variance do not depend on n. We do this again by centering and scaling, by setting $\boldsymbol{z}_n = \sqrt{n}(\overline{\boldsymbol{y}}_n - \boldsymbol{\mu})$. This has mean $\mathbf{0}$ and variance $\boldsymbol{V}$, which are independent of n as desired.

To develop a distributional approximation for $\boldsymbol{z}_n$ we use a multivariate central limit theorem. We present three such results, corresponding to the three univariate results from the previous section. Each is derived from the univariate theory by the Cramér-Wold device (Theorem 6.7.2).

We first present the multivariate version of Theorem 6.8.1.

Theorem 6.9.1 ***Multivariate Lindeberg–Lévy Central Limit Theorem.*** *If $\boldsymbol{y}_i \in \mathbb{R}^k$ are independent and identically distributed and $\mathbb{E}\|\boldsymbol{y}_i\|^2 < \infty$, then as $n \to \infty$*

$$\sqrt{n}(\overline{\boldsymbol{y}} - \boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{V})$$

where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y})$ and $\boldsymbol{V} = \mathbb{E}\left((\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'\right)$.

We next present a multivariate version of Theorem 6.8.2.

Theorem 6.9.2 ***Multivariate Lindeberg-Feller CLT.*** *Suppose $\boldsymbol{y}_{ni} \in \mathbb{R}^k$ are independent but not necessarily identically distributed with finite means $\boldsymbol{\mu}_{ni} = \mathbb{E}(\boldsymbol{y}_{ni})$ and variance matrices $\boldsymbol{V}_{ni} = \mathbb{E}\left((\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni})(\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni})'\right)$. Set $\overline{\boldsymbol{V}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{V}_{ni}$ and $\nu_n^2 = \lambda_{\min}(\overline{\boldsymbol{V}}_n)$. If $\nu_n^2 > 0$ and for all $\varepsilon > 0$*

$$\lim_{n\to\infty} \frac{1}{n\nu_n^2} \sum_{i=1}^n \mathbb{E}\left(\|\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\|^2 1\left(\|\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\|^2 \geq \varepsilon n \nu_n^2\right)\right) = 0 \tag{6.11}$$

then as $n \to \infty$

$$\overline{\boldsymbol{V}}_n^{-1/2}\sqrt{n}(\overline{\boldsymbol{y}} - \mathbb{E}(\overline{\boldsymbol{y}})) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{I}_k).$$

We finally present a multivariate version of Theorem 6.8.3.

Theorem 6.9.3 *Suppose $\boldsymbol{y}_{ni} \in \mathbb{R}^k$ are independent but not necessarily identically distributed with finite means $\boldsymbol{\mu}_{ni} = \mathbb{E}(\boldsymbol{y}_{ni})$ and variance matrices $\boldsymbol{V}_{ni} = \mathbb{E}\left((\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni})(\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni})'\right)$. Set $\overline{\boldsymbol{V}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{V}_{ni}$. If*

$$\overline{\boldsymbol{V}}_n \to \boldsymbol{V} > 0 \tag{6.12}$$

and for some $\delta > 0$

$$\sup_{n,i} \mathbb{E}\|\boldsymbol{y}_{ni}\|^{2+\delta} < \infty \tag{6.13}$$

then as $n \to \infty$

$$\sqrt{n}(\overline{\boldsymbol{y}} - \mathbb{E}(\overline{\boldsymbol{y}})) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{V}).$$

Similarly to Theorem 6.8.3, an advantage of Theorem 6.9.3 is that it allows the variance matrix $\boldsymbol{V}$ to be singular.

6.10 Higher Moments

Often we want to estimate a parameter $\boldsymbol{\mu}$ which is the expected value of a transformation of a random vector $\boldsymbol{y}$. That is, $\boldsymbol{\mu}$ can be written as

$$\boldsymbol{\mu} = \mathbb{E}\left(\boldsymbol{h}\left(\boldsymbol{y}\right)\right)$$

for some function $\boldsymbol{h} : \mathbb{R}^m \to \mathbb{R}^k$. For example, the second moment of y is $\mathbb{E}\left(y^2\right)$, the r^{th} is $\mathbb{E}\left(y^r\right)$, the moment generating function is $\mathbb{E}\left(\exp\left(ty\right)\right)$, and the distribution function is $\mathbb{E}\left(1\left\{y \leq x\right\}\right)$.

Estimating parameters of this form fits into our previous analysis by defining the random variable $\boldsymbol{z} = \boldsymbol{h}\left(\boldsymbol{y}\right)$ for then $\boldsymbol{\mu} = \mathbb{E}\left(\boldsymbol{z}\right)$ is just a simple moment of $\boldsymbol{z}$. This suggests the moment estimator

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}\left(\boldsymbol{y}_i\right).$$

For example, the moment estimator of $\mathbb{E}\left(y^r\right)$ is $n^{-1}\sum_{i=1}^{n} y_i^r$, that of the moment generating function is $n^{-1}\sum_{i=1}^{n} \exp\left(ty_i\right)$, and for the distribution function the estimator is $n^{-1}\sum_{i=1}^{n} 1\left\{y_i \leq x\right\}$.

Since $\widehat{\boldsymbol{\mu}}$ is a sample average, and transformations of iid variables are also i.i.d., the asymptotic results of the previous sections immediately apply.

Theorem 6.10.1 *If $\boldsymbol{y}_i$ are independent and identically distributed, $\boldsymbol{\mu} = \mathbb{E}\left(\boldsymbol{h}\left(\boldsymbol{y}\right)\right)$, and $\mathbb{E}\left\|\boldsymbol{h}\left(\boldsymbol{y}\right)\right\| < \infty$, then for $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}\left(\boldsymbol{y}_i\right)$, as $n \to \infty$, $\widehat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$.*

Theorem 6.10.2 *If $\boldsymbol{y}_i$ are independent and identically distributed, $\boldsymbol{\mu} = \mathbb{E}\left(\boldsymbol{h}\left(\boldsymbol{y}\right)\right)$, and $\mathbb{E}\left\|\boldsymbol{h}\left(\boldsymbol{y}\right)\right\|^2 < \infty$, then for $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}\left(\boldsymbol{y}_i\right)$, as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}\right)$$

where $\boldsymbol{V} = \mathbb{E}\left(\left(\boldsymbol{h}\left(\boldsymbol{y}\right) - \boldsymbol{\mu}\right)\left(\boldsymbol{h}\left(\boldsymbol{y}\right) - \boldsymbol{\mu}\right)'\right)$.

Theorems 6.10.1 and 6.10.2 show that the estimate $\widehat{\boldsymbol{\mu}}$ is consistent for $\boldsymbol{\mu}$ and asymptotically normally distributed, so long as the stated moment conditions hold.

A word of caution. Theorems 6.10.1 and 6.10.2 give the impression that it is possible to estimate any moment of y. Technically this is the case so long as that moment is finite. What is hidden by the notation, however, is that estimates of high order moments can be quite imprecise. For example, consider the sample 8^{th} moment $\widehat{\mu}_8 = \frac{1}{n}\sum_{i=1}^{n} y_i^8$, and suppose for simplicity that y is $\mathrm{N}(0,1)$. Then we can calculate[1] that $\mathrm{var}\left(\widehat{\mu}_8\right) = n^{-1}2,016,000$, which is immense, even for large n! In general, higher-order moments are challenging to estimate because their variance depends upon even higher moments which can be quite large in some cases.

[1] By the formula for the variance of a mean $\mathrm{var}\left(\widehat{\mu}_8\right) = n^{-1}\left(\mathbb{E}\left(y^{16}\right) - \left(\mathbb{E}\left(y^8\right)\right)^2\right)$. Since y is $\mathrm{N}(0,1)$, $\mathbb{E}\left(y^{16}\right) = 15!! = 2,027,025$ and $\mathbb{E}\left(y^8\right) = 7!! = 105$ where $k!!$ is the double factorial.

6.11 Functions of Moments

We now expand our investigation and consider estimation of parameters which can be written as a continuous function of $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{h}(\boldsymbol{y}))$. That is, the parameter of interest can be written as

$$\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{g}(\mathbb{E}(\boldsymbol{h}(\boldsymbol{y}))) \tag{6.14}$$

for some functions $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^\ell$ and $\boldsymbol{h} : \mathbb{R}^m \to \mathbb{R}^k$.

As one example, the geometric mean of wages w is

$$\gamma = \exp(\mathbb{E}(\log(w))). \tag{6.15}$$

This is (6.14) with $g(u) = \exp(u)$ and $h(w) = \log(w)$.

A simple yet common example is the variance

$$\begin{aligned}\sigma^2 &= \mathbb{E}(w - \mathbb{E}(w))^2 \\ &= \mathbb{E}(w^2) - (\mathbb{E}(w))^2.\end{aligned}$$

This is (6.14) with

$$\boldsymbol{h}(w) = \begin{pmatrix} w \\ w^2 \end{pmatrix}$$

and

$$g(\mu_1, \mu_2) = \mu_2 - \mu_1^2.$$

Similarly, the skewness of the wage distribution is

$$sk = \frac{\mathbb{E}\left((w - \mathbb{E}(w))^3\right)}{\left(\mathbb{E}\left((w - \mathbb{E}(w))^2\right)\right)^{3/2}}.$$

This is (6.14) with

$$\boldsymbol{h}(w) = \begin{pmatrix} w \\ w^2 \\ w^3 \end{pmatrix}$$

and

$$g(\mu_1, \mu_2, \mu_3) = \frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{\left(\mu_2 - \mu_1^2\right)^{3/2}}. \tag{6.16}$$

The parameter $\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\mu})$ is not a population moment, so it does not have a direct moment estimator. Instead, it is common to use a **plug-in estimate** formed by replacing the unknown $\boldsymbol{\mu}$ with its point estimate $\widehat{\boldsymbol{\mu}}$ and then "plugging" this into the expression for $\boldsymbol{\beta}$. The first step is

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i)$$

and the second step is

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}}).$$

Again, the hat "^" indicates that $\widehat{\boldsymbol{\beta}}$ is a sample estimate of $\boldsymbol{\beta}$.

For example, the plug-in estimate of the geometric mean γ of the wage distribution from (6.15) is

$$\widehat{\gamma} = \exp(\widehat{\mu})$$

with

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \log(wage_i).$$

The plug-in estimate of the variance is

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n} w_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} w_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} (w_i - \overline{w})^2 .\end{aligned}$$

The estimator for the skewness is

$$\begin{aligned}\widehat{sk} &= \frac{\widehat{\mu}_3 - 3\widehat{\mu}_2\widehat{\mu}_1 + 2\widehat{\mu}_1^3}{\left(\widehat{\mu}_2 - \widehat{\mu}_1^2\right)^{3/2}} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} (w_i - \overline{w})^3}{\left(\frac{1}{n}\sum_{i=1}^{n} (w_i - \overline{w})^2\right)^{3/2}}\end{aligned}$$

where

$$\widehat{\mu}_j = \frac{1}{n}\sum_{i=1}^{n} w_i^j.$$

A useful property is that continuous functions are limit-preserving.

Theorem 6.11.1 ***Continuous Mapping Theorem (CMT).*** *If $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{c}$ as $n \to \infty$ and $\boldsymbol{g}(\cdot)$ is continuous at $\boldsymbol{c}$, then $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{p} \boldsymbol{g}(\boldsymbol{c})$ as $n \to \infty$.*

The proof of Theorem 6.11.1 is given in Section 6.16.

For example, if $z_n \xrightarrow{p} c$ as $n \to \infty$ then

$$\begin{aligned} z_n + a &\xrightarrow{p} c + a \\ az_n &\xrightarrow{p} ac \\ z_n^2 &\xrightarrow{p} c^2 \end{aligned}$$

as the functions $g(u) = u + a$, $g(u) = au$, and $g(u) = u^2$ are continuous. Also

$$\frac{a}{z_n} \xrightarrow{p} \frac{a}{c}$$

if $c \neq 0$. The condition $c \neq 0$ is important as the function $g(u) = a/u$ is not continuous at $u = 0$.

If $\boldsymbol{y}_i$ are independent and identically distributed, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{h}(\boldsymbol{y}))$, and $\mathbb{E}\|\boldsymbol{h}(\boldsymbol{y})\| < \infty$, then for $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i)$, as $n \to \infty$, $\widehat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$. Applying the CMT, $\widehat{\boldsymbol{\beta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}}) \xrightarrow{p} \boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{\beta}$.

Theorem 6.11.2 *If $\boldsymbol{y}_i$ are independent and identically distributed, $\boldsymbol{\beta} = \boldsymbol{g}(\mathbb{E}(\boldsymbol{h}(\boldsymbol{y}))), \mathbb{E}\|\boldsymbol{h}(\boldsymbol{y})\| < \infty$, and $\boldsymbol{g}(\boldsymbol{u})$ is continuous at $\boldsymbol{u} = \boldsymbol{\mu}$, then for $\widehat{\boldsymbol{\beta}} = \boldsymbol{g}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i)\right)$, as $n \to \infty$, $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.*

To apply Theorem 6.11.2 it is necessary to check if the function $\boldsymbol{g}$ is continuous at $\boldsymbol{\mu}$. In our first example $g(u) = \exp(u)$ is continuous everywhere. It therefore follows from Theorem 6.6.2 and Theorem 6.11.2 that if $\mathbb{E}|\log(wage)| < \infty$ then as $n \to \infty$, $\widehat{\gamma} \xrightarrow{p} \gamma$.

In the example of the variance, g is continuous for all $\boldsymbol{\mu}$. Thus if $\mathbb{E}\left(w^2\right) < \infty$ then as $n \to \infty$, $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$.

In our third example g defined in (6.16) is continuous for all $\boldsymbol{\mu}$ such that $\text{var}(w) = \mu_2 - \mu_1^2 > 0$, which holds unless w has a degenerate distribution. Thus if $\mathbb{E}|w|^3 < \infty$ and $\text{var}(w) > 0$ then as $n \to \infty$, $\widehat{sk} \xrightarrow{p} sk$.

6.12 Delta Method

In this section we introduce two tools – an extended version of the CMT and the Delta Method – which allow us to calculate the asymptotic distribution of the parameter estimate $\widehat{\boldsymbol{\beta}}$.

We first present an extended version of the continuous mapping theorem which allows convergence in distribution.

Theorem 6.12.1 ***Continuous Mapping Theorem***
If $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ *as* $n \to \infty$ *and* $\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^k$ *has the set of discontinuity points* D_g *such that* $\Pr(\boldsymbol{z} \in D_g) = 0$*, then* $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{d} \boldsymbol{g}(\boldsymbol{z})$ *as* $n \to \infty$*.*

For a proof of Theorem 6.12.1 see Theorem 2.3 of van der Vaart (1998). It was first proved by Mann and Wald (1943) and is therefore sometimes referred to as the Mann-Wald Theorem.

Theorem 6.12.1 allows the function $\boldsymbol{g}$ to be discontinuous only if the probability at being at a discontinuity point is zero. For example, the function $g(u) = u^{-1}$ is discontinuous at $u = 0$, but if $z_n \xrightarrow{d} z \sim \mathrm{N}(0, 1)$ then $\Pr(z = 0) = 0$ so $z_n^{-1} \xrightarrow{d} z^{-1}$.

A special case of the Continuous Mapping Theorem is known as Slutsky's Theorem.

Theorem 6.12.2 ***Slutsky's Theorem***
If $z_n \xrightarrow{d} z$ *and* $c_n \xrightarrow{p} c$ *as* $n \to \infty$*, then*

1. $z_n + c_n \xrightarrow{d} z + c$
2. $z_n c_n \xrightarrow{d} zc$
3. $\dfrac{z_n}{c_n} \xrightarrow{d} \dfrac{z}{c}$ *if* $c \neq 0$

Even though Slutsky's Theorem is a special case of the CMT, it is a useful statement as it focuses on the most common applications – addition, multiplication, and division.

Despite the fact that the plug-in estimator $\widehat{\boldsymbol{\beta}}$ is a function of $\widehat{\boldsymbol{\mu}}$ for which we have an asymptotic distribution, Theorem 6.12.1 does not directly give us an asymptotic distribution for $\widehat{\boldsymbol{\beta}}$. This is because $\widehat{\boldsymbol{\beta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}})$ is written as a function of $\widehat{\boldsymbol{\mu}}$, not of the standardized sequence $\sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})$. We need an intermediate step – a first order Taylor series expansion. This step is so critical to statistical theory that it has its own name – **The Delta Method**.

Theorem 6.12.3 ***Delta Method:***
If $\sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} \boldsymbol{\xi}$*, where* $\boldsymbol{g}(\boldsymbol{u})$ *is continuously differentiable in a neighborhood of* $\boldsymbol{\mu}$ *then as* $n \to \infty$

$$\sqrt{n}(\boldsymbol{g}(\widehat{\boldsymbol{\mu}}) - \boldsymbol{g}(\boldsymbol{\mu})) \xrightarrow{d} \boldsymbol{G}'\boldsymbol{\xi} \tag{6.17}$$

where $\boldsymbol{G}(\boldsymbol{u}) = \frac{\partial}{\partial \boldsymbol{u}} \boldsymbol{g}(\boldsymbol{u})'$ *and* $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\mu})$*. In particular, if* $\boldsymbol{\xi} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{V})$ *then as* $n \to \infty$

$$\sqrt{n}(\boldsymbol{g}(\widehat{\boldsymbol{\mu}}) - \boldsymbol{g}(\boldsymbol{\mu})) \xrightarrow{d} \mathrm{N}\left(0, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}\right). \tag{6.18}$$

The Delta Method allows us to complete our derivation of the asymptotic distribution of the estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. By combining Theorems 6.10.2 and 6.12.3 we can find the asymptotic distribution of the plug-in estimator $\widehat{\boldsymbol{\beta}}$.

Theorem 6.12.4 *If $\boldsymbol{y}_i$ are independent and identically distributed, $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{h}(\boldsymbol{y}))$, $\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\mu})$, $\mathbb{E}\|\boldsymbol{h}(\boldsymbol{y})\|^2 < \infty$, and $\boldsymbol{G}(\boldsymbol{u}) = \dfrac{\partial}{\partial \boldsymbol{u}} \boldsymbol{g}(\boldsymbol{u})'$ is continuous in a neighborhood of $\boldsymbol{\mu}$, then for $\widehat{\boldsymbol{\beta}} = \boldsymbol{g}\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{h}(\boldsymbol{y}_i)\right)$, as $n \to \infty$*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}\right)$$

where $\boldsymbol{V} = \mathbb{E}\left((\boldsymbol{h}(\boldsymbol{y}) - \boldsymbol{\mu})(\boldsymbol{h}(\boldsymbol{y}) - \boldsymbol{\mu})'\right)$ and $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\mu})$.

Theorem 6.11.2 established the consistency of $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, and Theorem 6.12.4 established its asymptotic normality. It is instructive to compare the conditions required for these results. Consistency required that $\boldsymbol{h}(\boldsymbol{y})$ have a finite mean, while asymptotic normality requires that this variable have a finite variance. Consistency required that $\boldsymbol{g}(\boldsymbol{u})$ be continuous, while our proof of asymptotic normality used the assumption that $\boldsymbol{g}(\boldsymbol{u})$ is continuously differentiable.

6.13 Stochastic Order Symbols

It is convenient to have simple symbols for random variables and vectors which converge in probability to zero or are stochastically bounded. In this section we introduce some of the most commonly found notation.

It might be useful to review the common notation for non-random convergence and boundedness. Let x_n and a_n, $n = 1, 2, ...$, be non-random sequences. The notation

$$x_n = o(1)$$

(pronounced "small oh-one") is equivalent to $x_n \to 0$ as $n \to \infty$. The notation

$$x_n = o(a_n)$$

is equivalent to $a_n^{-1} x_n \to 0$ as $n \to \infty$. The notation

$$x_n = O(1)$$

(pronounced "big oh-one") means that x_n is bounded uniformly in n – there exists an $M < \infty$ such that $|x_n| \leq M$ for all n. The notation

$$x_n = O(a_n)$$

is equivalent to $a_n^{-1} x_n = O(1)$.

We now introduce similar concepts for sequences of random variables. Let z_n and a_n, $n = 1, 2, ...$ be sequences of random variables. (In most applications, a_n is non-random.) The notation

$$z_n = o_p(1)$$

("small oh-P-one") means that $z_n \xrightarrow{p} 0$ as $n \to \infty$. For example, for any consistent estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ we can write

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_p(1).$$

We also write

$$z_n = o_p(a_n)$$

if $a_n^{-1} z_n = o_p(1)$.

Similarly, the notation $z_n = O_p(1)$ ("big oh-P-one") means that z_n is bounded in probability. Precisely, for any $\varepsilon > 0$ there is a constant $M_\varepsilon < \infty$ such that

$$\limsup_{n\to\infty} \Pr\left(|z_n| > M_\varepsilon\right) \leq \varepsilon.$$

Furthermore, we write

$$z_n = O_p(a_n)$$

if $a_n^{-1} z_n = O_p(1)$.

$O_p(1)$ is weaker than $o_p(1)$ in the sense that $z_n = o_p(1)$ implies $z_n = O_p(1)$ but not the reverse. However, if $z_n = O_p(a_n)$ then $z_n = o_p(b_n)$ for any b_n such that $a_n/b_n \to 0$.

If a random vector converges in distribution $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ (for example, if $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{V})$) then $\boldsymbol{z}_n = O_p(1)$. It follows that for estimators $\widehat{\boldsymbol{\beta}}$ which satisfy the convergence of Theorem 6.12.4 then we can write

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + O_p(n^{-1/2}).$$

In words, this statement says that the estimator $\widehat{\boldsymbol{\beta}}$ equals the true coefficient $\boldsymbol{\beta}$ plus a random component which is bounded when scaled by $n^{1/2}$. Equivalently, we can write

$$n^{1/2}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = O_p(1).$$

Another useful observation is that a random sequence with a bounded moment is stochastically bounded.

Theorem 6.13.1 *If $\boldsymbol{z}_n$ is a random vector which satisfies*

$$\mathbb{E}\left\|\boldsymbol{z}_n\right\|^\delta = O\left(a_n\right)$$

for some sequence a_n and $\delta > 0$, then

$$\boldsymbol{z}_n = O_p(a_n^{1/\delta}).$$

Similarly, $\mathbb{E}\left\|\boldsymbol{z}_n\right\|^\delta = o\left(a_n\right)$ implies $\boldsymbol{z}_n = o_p(a_n^{1/\delta})$.

This can be shown using Markov's inequality (B.14). The assumptions imply that there is some $M < \infty$ such that $\mathbb{E}\left\|\boldsymbol{z}_n\right\|^\delta \leq M a_n$ for all n. For any ε set $B = \left(\frac{M}{\varepsilon}\right)^{1/\delta}$. Then

$$\Pr\left(a_n^{-1/\delta}\left\|\boldsymbol{z}_n\right\| > B\right) = \Pr\left(\left\|\boldsymbol{z}_n\right\|^\delta > \frac{M a_n}{\varepsilon}\right) \leq \frac{\varepsilon}{M a_n}\mathbb{E}\left\|\boldsymbol{z}_n\right\|^\delta \leq \varepsilon$$

as required.

There are many simple rules for manipulating $o_p(1)$ and $O_p(1)$ sequences which can be deduced from the continuous mapping theorem or Slutsky's Theorem. For example,

$$\begin{aligned}
o_p(1) + o_p(1) &= o_p(1) \\
o_p(1) + O_p(1) &= O_p(1) \\
O_p(1) + O_p(1) &= O_p(1) \\
o_p(1)o_p(1) &= o_p(1) \\
o_p(1)O_p(1) &= o_p(1) \\
O_p(1)O_p(1) &= O_p(1).
\end{aligned}$$

6.14 Uniform Stochastic Bounds*

For some applications it can be useful to obtain the stochastic order of the random variable

$$\max_{1\leq i\leq n}|y_i|.$$

This is the magnitude of the largest observation in the sample $\{y_1,...,y_n\}$. If the support of the distribution of y_i is unbounded, then as the sample size n increases, the largest observation will also tend to increase. It turns out that there is a simple characterization.

Theorem 6.14.1 *If $\boldsymbol{y}_i$ are identically distributed and $\mathbb{E}|y|^r<\infty$, then as $n\to\infty$*

$$n^{-1/r}\max_{1\leq i\leq n}|y_i|\xrightarrow{p}0. \tag{6.19}$$

Furthermore, if $\mathbb{E}(\exp(ty))<\infty$ for some $t>0$ then for any $\eta>0$

$$(\log n)^{-(1+\eta)}\max_{1\leq i\leq n}|y_i|\xrightarrow{p}0. \tag{6.20}$$

The proof of Theorem 6.14.1 is presented in Section 6.16.

Equivalently, (6.19) can be written as

$$\max_{1\leq i\leq n}|y_i|=o_p(n^{1/r}) \tag{6.21}$$

and (6.22) as

$$\max_{1\leq i\leq n}|y_i|=o_p(\log n). \tag{6.22}$$

Equation (6.21) says that if y has r finite moments, then the largest observation will diverge at a rate slower than $n^{1/r}$. As r increases this rate decreases. Equation (6.22) shows that if we strengthen this to y having all finite moments and a finite moment generating function (for example, if y is normally distributed) then the largest observation will diverge slower than $\log n$. Thus the higher the moments, the slower the rate of divergence.

To simplify the notation, we write (6.21) as $y_i=o_p(n^{1/r})$ uniformly in $1\leq i\leq n$. It is important to understand when the O_p or o_p symbols are applied to subscript i random variables whether the convergence is pointwise in i, or is uniform in i in the sense of (6.21)-(6.22).

Theorem 6.14.1 applies to random vectors. For example, if $\mathbb{E}\|\boldsymbol{y}\|^r<\infty$ then

$$\max_{1\leq i\leq n}\|\boldsymbol{y}_i\|=o_p(n^{1/r}).$$

6.15 Semiparametric Efficiency

In this section we argue that the sample mean $\widehat{\boldsymbol{\mu}}$ and plug-in estimator $\widehat{\boldsymbol{\beta}}=\mathbf{g}(\widehat{\boldsymbol{\mu}})$ are efficient estimators of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\beta}$. Our demonstration is based on the rich but technically challenging theory of semiparametric efficiency bounds. An excellent accessible review has been provided by Newey (1990). We will also appeal to the asymptotic theory of maximum likelihood estimation (see Chapter 5).

We start by examining the sample mean $\widehat{\boldsymbol{\mu}}$, for the asymptotic efficiency of $\widehat{\boldsymbol{\beta}}$ will follow from that of $\widehat{\boldsymbol{\mu}}$.

Recall, we know that if $\mathbb{E}\left(\|\boldsymbol{y}\|^2\right)<\infty$ then the sample mean has the asymptotic distribution $\sqrt{n}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})\xrightarrow{d}\mathrm{N}(0,\boldsymbol{V})$. We want to know if $\widehat{\boldsymbol{\mu}}$ is the best feasible estimator, or if there is another

estimator with a smaller asymptotic variance. While it seems intuitively unlikely that another estimator could have a smaller asymptotic variance, how do we know that this is not the case?

When we ask if $\widehat{\boldsymbol{\mu}}$ is the best estimator, we need to be clear about the class of models – the class of permissible distributions. For estimation of the mean $\boldsymbol{\mu}$ of the distribution of $\boldsymbol{y}$ the broadest conceivable class is $\mathcal{L}_1 = \{F : \mathbb{E}\|\boldsymbol{y}\| < \infty\}$. This class is too broad for our current purposes, as $\widehat{\boldsymbol{\mu}}$ is not asymptotically $\mathrm{N}(0, \boldsymbol{V})$ for all $F \in \mathcal{L}_1$. A more realistic choice is $\mathcal{L}_2 = \left\{F : \mathbb{E}\left(\|\boldsymbol{y}\|^2\right) < \infty\right\}$ – the class of finite-variance distributions. When we seek an efficient estimator of the mean $\boldsymbol{\mu}$ in the class of models $\mathcal{L}_2$ what we are seeking is the best estimator, given that all we know is that $F \in \mathcal{L}_2$.

To show that the answer is not immediately obvious, it might be helpful to review a setting where the sample mean is inefficient. Suppose that $y \in \mathbb{R}$ has the double exponential density $f(y \mid \mu) = 2^{-1/2}\exp\left(-|y-\mu|\sqrt{2}\right)$. Since $\mathrm{var}(y) = 1$ we see that the sample mean satisfies $\sqrt{n}(\widetilde{\mu} - \mu) \xrightarrow{d} \mathrm{N}(0,1)$. In this model the maximum likelihood estimator (MLE) $\widetilde{\mu}$ for μ is the sample median. Recall from the theory of maximum likelihood that the MLE satisfies $\sqrt{n}(\widetilde{\mu} - \mu) \xrightarrow{d} \mathrm{N}\left(0, \left(\mathbb{E}\left(S^2\right)\right)^{-1}\right)$ where $S = \frac{\partial}{\partial\mu}\log f(y \mid \mu) = -\sqrt{2}\,\mathrm{sgn}(y-\mu)$ is the score. We can calculate that $\mathbb{E}\left(S^2\right) = 2$ and thus conclude that $\sqrt{n}(\widetilde{\mu} - \mu) \xrightarrow{d} \mathrm{N}(0, 1/2)$. The asymptotic variance of the MLE is one-half that of the sample mean. Thus when the true density is known to be double exponential the sample mean is inefficient.

But the estimator which achieves this improved efficiency – the sample median – is not generically consistent for the population mean. It is inconsistent if the density is asymmetric or skewed. So the improvement comes at a great cost. Another way of looking at this is that the sample median is efficient in the class of densities $\left\{f(y \mid \mu) = 2^{-1/2}\exp\left(-|y-\mu|\sqrt{2}\right)\right\}$ but unless it is known that this is the correct distribution class this knowledge is not very useful.

The relevant question is whether or not the sample mean is efficient when the form of the distribution is unknown. We call this setting **semiparametric** as the parameter of interest (the mean) is finite dimensional while the remaining features of the distribution are unspecified. In the semiparametric context an estimator is called **semiparametrically efficient** if it has the smallest asymptotic variance among all semiparametric estimators.

The mathematical trick is to reduce the semiparametric model to a set of parametric "submodels". The Cramer-Rao variance bound can be found for each parametric submodel. The variance bound for the semiparametric model (the union of the submodels) is then defined as the supremum of the individual variance bounds.

Formally, suppose that the true density of $\boldsymbol{y}$ is the unknown function $f(\boldsymbol{y})$ with mean $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{y}) = \int \boldsymbol{y} f(\boldsymbol{y}) d\boldsymbol{y}$. A parametric submodel η for $f(\boldsymbol{y})$ is a density $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta})$ which is a smooth function of a parameter $\boldsymbol{\theta}$, and there is a true value $\boldsymbol{\theta}_0$ such that $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta}_0) = f(\boldsymbol{y})$. The index η indicates the submodels. The equality $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta}_0) = f(\boldsymbol{y})$ means that the submodel class passes through the true density, so the submodel is a true model. The class of submodels η and parameter $\boldsymbol{\theta}_0$ depend on the true density f. In the submodel $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta})$, the mean is $\boldsymbol{\mu}_\eta(\boldsymbol{\theta}) = \int \boldsymbol{y} f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta})\, d\boldsymbol{y}$ which varies with the parameter $\boldsymbol{\theta}$. Let $\eta \in \aleph$ be the class of all submodels for f.

Since each submodel η is parametric we can calculate the efficiency bound for estimation of $\boldsymbol{\mu}$ within this submodel. Specifically, given the density $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta})$ its likelihood score is

$$\boldsymbol{S}_\eta = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta}_0),$$

so the Cramer-Rao lower bound for estimation of $\boldsymbol{\theta}$ is $\left(\mathbb{E}\left(\boldsymbol{S}_{\boldsymbol{\eta}}\boldsymbol{S}_{\boldsymbol{\eta}}'\right)\right)^{-1}$. Defining $\boldsymbol{M}_\eta = \frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\mu}_\eta(\boldsymbol{\theta}_0)'$, by Theorem 5.16.3 the Cramer-Rao lower bound for estimation of $\boldsymbol{\mu}$ within the submodel η is $\boldsymbol{V}_\eta = \boldsymbol{M}_\eta'\left(\mathbb{E}\left(\boldsymbol{S}_{\boldsymbol{\eta}}\boldsymbol{S}_{\boldsymbol{\eta}}'\right)\right)^{-1}\boldsymbol{M}_\eta$.

As $\boldsymbol{V}_\eta$ is the efficiency bound for the submodel class $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta})$, no estimator can have an asymptotic variance smaller than $\boldsymbol{V}_\eta$ for any density $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta})$ in the submodel class, including the

true density f. This is true for all submodels η. Thus the asymptotic variance of any semiparametric estimator cannot be smaller than $\boldsymbol{V}_\eta$ for any conceivable submodel. Taking the supremum of the Cramer-Rao bounds from all conceivable submodels we define[2]

$$\overline{\boldsymbol{V}} = \sup_{\eta \in \aleph} \boldsymbol{V}_\eta.$$

The asymptotic variance of any semiparametric estimator cannot be smaller than $\overline{\boldsymbol{V}}$, since it cannot be smaller than any individual $\boldsymbol{V}_\eta$. We call $\overline{\boldsymbol{V}}$ the **semiparametric asymptotic variance bound** or **semiparametric efficiency bound** for estimation of $\boldsymbol{\mu}$, as it is a lower bound on the asymptotic variance for any semiparametric estimator. If the asymptotic variance of a specific semiparametric estimator equals the bound $\overline{\boldsymbol{V}}$ we say that the estimator is **semiparametrically efficient**.

For many statistical problems it is quite challenging to calculate the semiparametric variance bound. However, in some cases there is a simple method to find the solution. Suppose that we can find a submodel η_0 whose Cramer-Rao lower bound satisfies $\boldsymbol{V}_{\eta_0} = \boldsymbol{V}_{\boldsymbol{\mu}}$ where $\boldsymbol{V}_{\boldsymbol{\mu}}$ is the asymptotic variance of a known semiparametric estimator. In this case, we can deduce that $\overline{\boldsymbol{V}} = \boldsymbol{V}_{\eta_0} = \boldsymbol{V}_{\boldsymbol{\mu}}$. Otherwise (that is, if $\boldsymbol{V}_{\eta_0}$ is not the efficiency bound) there would exist another submodel η_1 whose Cramer-Rao lower bound satisfies $\boldsymbol{V}_{\eta_0} < \boldsymbol{V}_{\eta_1}$ (because $\boldsymbol{V}_{\eta_0}$ is not the supremum). This would imply $\boldsymbol{V}_{\boldsymbol{\mu}} < \boldsymbol{V}_{\eta_1}$ which contradicts the Cramer-Rao Theorem (since when submodel η_1 is true then no estimator can have a lower variance than $\boldsymbol{V}_{\eta_1}$).

We now find this submodel for the sample mean $\widehat{\boldsymbol{\mu}}$. Our goal is to find a parametric submodel whose Cramer-Rao bound for $\boldsymbol{\mu}$ is $\boldsymbol{V}$. This can be done by creating a tilted version of the true density. Consider the parametric submodel

$$f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta}) = f(\boldsymbol{y})\left(1 + \boldsymbol{\theta}' \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right) \tag{6.23}$$

where $f(\boldsymbol{y})$ is the true density and $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{y}$. Note that

$$\int f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta})\, d\boldsymbol{y} = \int f(\boldsymbol{y}) d\boldsymbol{y} + \boldsymbol{\theta}' \boldsymbol{V}^{-1} \int f(\boldsymbol{y})(\boldsymbol{y} - \boldsymbol{\mu})\, d\boldsymbol{y} = 1$$

and for all $\boldsymbol{\theta}$ close to zero $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta}) \geq 0$. Thus $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta})$ is a valid density function. It is a parametric submodel since $f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta}_0) = f(\boldsymbol{y})$ when $\boldsymbol{\theta}_0 = 0$. This parametric submodel has the mean

$$\begin{aligned}
\boldsymbol{\mu}(\boldsymbol{\theta}) &= \int \boldsymbol{y} f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta})\, d\boldsymbol{y} \\
&= \int \boldsymbol{y} f(\boldsymbol{y}) d\boldsymbol{y} + \int f(\boldsymbol{y}) \boldsymbol{y} (\boldsymbol{y} - \boldsymbol{\mu})' \boldsymbol{V}^{-1} \boldsymbol{\theta} d\boldsymbol{y} \\
&= \boldsymbol{\mu} + \boldsymbol{\theta}
\end{aligned}$$

which is a smooth function of $\boldsymbol{\theta}$.

Since

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(1 + \boldsymbol{\theta}' \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right) = \frac{\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})}{1 + \boldsymbol{\theta}' \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})}$$

it follows that the score function for $\boldsymbol{\theta}$ is

$$\boldsymbol{S}_\eta = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta}_0) = \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}). \tag{6.24}$$

By Theorem 5.16.3 the Cramer-Rao lower bound for $\boldsymbol{\theta}$ is

$$\left(\mathbb{E}\left(\boldsymbol{S}_\eta \boldsymbol{S}_\eta'\right)\right)^{-1} = \left(\boldsymbol{V}^{-1} \mathbb{E}\left((\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'\right) \boldsymbol{V}^{-1}\right)^{-1} = \boldsymbol{V}. \tag{6.25}$$

[2]It is not obvious that this supremum exists, as $\boldsymbol{V}_\eta$ is a matrix so there is not a unique ordering of matrices. However, in many cases (including the ones we study) the supremum exists and is unique.

The Cramer-Rao lower bound for $\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\mu} + \boldsymbol{\theta}$ is also $\boldsymbol{V}$, and this equals the asymptotic variance of the moment estimator $\widehat{\boldsymbol{\mu}}$. This was what we set out to show.

In summary, we have shown that in the submodel (6.23) the Cramer-Rao lower bound for estimation of $\boldsymbol{\mu}$ is $\boldsymbol{V}$ which equals the asymptotic variance of the sample mean. This establishes the following result.

Proposition 6.15.1 *In the class of distributions $F \in \mathcal{L}_2$, the semiparametric variance bound for estimation of $\boldsymbol{\mu}$ is $\boldsymbol{V} = \mathrm{var}(y_i)$, and the sample mean $\widehat{\boldsymbol{\mu}}$ is a semiparametrically efficient estimator of the population mean $\boldsymbol{\mu}$.*

We call this result a proposition rather than a theorem as we have not attended to the regularity conditions.

It is a simple matter to extend this result to the plug-in estimator $\widehat{\boldsymbol{\beta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}})$. We know from Theorem 6.12.4 that if $\mathbb{E}\|\boldsymbol{y}\|^2 < \infty$ and $\boldsymbol{g}(\boldsymbol{u})$ is continuously differentiable at $\boldsymbol{u} = \boldsymbol{\mu}$ then the plug-in estimator has the asymptotic distribution $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G})$. We therefore consider the class of distributions

$$\mathcal{L}_2(\boldsymbol{g}) = \left\{F : \mathbb{E}\|\boldsymbol{y}\|^2 < \infty,\ \boldsymbol{g}(\boldsymbol{u}) \text{ is continuously differentiable at } \boldsymbol{u} = \mathbb{E}(\boldsymbol{y})\right\}.$$

For example, if $\beta = \mu_1/\mu_2$ where $\mu_1 = \mathbb{E}(y_1)$ and $\mu_2 = \mathbb{E}(y_2)$ then

$$\mathcal{L}_2(g) = \left\{F : \mathbb{E}\left(y_1^2\right) < \infty,\ \mathbb{E}\left(y_2^2\right) < \infty, \text{ and } \mathbb{E}(y_2) \neq 0\right\}.$$

For any submodel η the Cramer-Rao lower bound for estimation of $\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\mu})$ is $\boldsymbol{G}'\boldsymbol{V}_\eta\boldsymbol{G}$. For the submodel (6.23) this bound is $\boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}$ which equals the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ from Theorem 6.12.4. Thus $\widehat{\boldsymbol{\beta}}$ is semiparametrically efficient.

Proposition 6.15.2 *In the class of distributions $F \in \mathcal{L}_2(\boldsymbol{g})$ the semiparametric variance bound for estimation of $\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\mu})$ is $\boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}$, and the plug-in estimator $\widehat{\boldsymbol{\beta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}})$ is a semiparametrically efficient estimator of $\boldsymbol{\beta}$.*

The result in Proposition 6.15.2 is quite general. Smooth functions of sample moments are efficient estimators for their population counterparts. This is a very powerful result, as most econometric estimators can be written (or approximated) as smooth functions of sample means.

6.16 Technical Proofs*

In this section we provide proofs of some of the more technical points in the chapter. These proofs may only be of interest to more mathematically inclined students.

Proof of Theorem 6.4.2: Without loss of generality, we can assume $\mathbb{E}(y_i) = 0$ by recentering y_i on its expectation.

We need to show that for all $\delta > 0$ and $\eta > 0$ there is some $N < \infty$ so that for all $n \geq N$, $\Pr(|\overline{y}| > \delta) \leq \eta$. Fix δ and η. Set $\varepsilon = \delta\eta/3$. Pick $C < \infty$ large enough so that

$$\mathbb{E}(|y_i| 1(|y_i| > C)) \leq \varepsilon \tag{6.26}$$

(where $1(\cdot)$ is the indicator function) which is possible since $\mathbb{E}|y_i| < \infty$. Define the random variables

$$w_i = y_i 1(|y_i| \le C) - \mathbb{E}(y_i 1(|y_i| \le C))$$
$$z_i = y_i 1(|y_i| > C) - \mathbb{E}(y_i 1(|y_i| > C))$$

so that

$$\overline{y} = \overline{w} + \overline{z}$$

and

$$\mathbb{E}|\overline{y}| \le \mathbb{E}|\overline{w}| + \mathbb{E}|\overline{z}|. \tag{6.27}$$

We now show that sum of the expectations on the right-hand-side can be bounded below 3ε.

First, by the Triangle Inequality (A.26) and the Expectation Inequality (B.8),

$$\begin{aligned}\mathbb{E}|z_i| &= \mathbb{E}|y_i 1(|y_i| > C) - \mathbb{E}(y_i 1(|y_i| > C))| \\ &\le \mathbb{E}|y_i 1(|y_i| > C)| + |\mathbb{E}(y_i 1(|y_i| > C))| \\ &\le 2\mathbb{E}|y_i 1(|y_i| > C)| \\ &\le 2\varepsilon, \end{aligned} \tag{6.28}$$

and thus by the Triangle Inequality (A.26) and (6.28)

$$\mathbb{E}|\overline{z}| = \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} z_i\right| \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|z_i| \le 2\varepsilon. \tag{6.29}$$

Second, by a similar argument

$$\begin{aligned}|w_i| &= |y_i 1(|y_i| \le C) - \mathbb{E}(y_i 1(|y_i| \le C))| \\ &\le |y_i 1(|y_i| \le C)| + |\mathbb{E}(y_i 1(|y_i| \le C))| \\ &\le 2|y_i 1(|y_i| \le C)| \\ &\le 2C \end{aligned} \tag{6.30}$$

where the final inequality is (6.26). Then by Jensen's Inequality (B.5), the fact that the w_i are iid and mean zero, and (6.30),

$$(\mathbb{E}|\overline{w}|)^2 \le \mathbb{E}\left(|\overline{w}|^2\right) = \frac{\mathbb{E}\left(w_i^2\right)}{n} \le \frac{4C^2}{n} \le \varepsilon^2 \tag{6.31}$$

the final inequality holding for $n \ge 4C^2/\varepsilon^2 = 36C^2/\delta^2\eta^2$. Equations (6.27), (6.29) and (6.31) together show that

$$\mathbb{E}|\overline{y}| \le 3\varepsilon \tag{6.32}$$

as desired.

Finally, by Markov's Inequality (B.14) and (6.32),

$$\Pr(|\overline{y}| > \delta) \le \frac{\mathbb{E}|\overline{y}|}{\delta} \le \frac{3\varepsilon}{\delta} = \eta,$$

the final equality by the definition of ε. We have shown that for any $\delta > 0$ and $\eta > 0$ then for all $n \ge 36C^2/\delta^2\eta^2$, $\Pr(|\overline{y}| > \delta) \le \eta$, as needed. ■

Proof of Theorem 6.6.1: By Loève's c_r Inequality (A.16)

$$\|\boldsymbol{y}\| = \left(\sum_{j=1}^{m} y_j^2\right)^{1/2} \le \sum_{j=1}^{m}|y_j|.$$

Thus if $\mathbb{E}|y_j| < \infty$ for $j = 1, ..., m$, then

$$\mathbb{E}\|\boldsymbol{y}\| \leq \sum_{j=1}^{m} \mathbb{E}|y_j| < \infty.$$

For the reverse inequality, the Euclidean norm of a vector is larger than the length of any individual component, so for any j, $|y_j| \leq \|\boldsymbol{y}\|$. Thus, if $\mathbb{E}\|\boldsymbol{y}\| < \infty$, then $\mathbb{E}|y_j| < \infty$ for $j = 1, ..., m$. ■

Proof of Theorem 6.7.2: By Lévy's Continuity Theorem (Theorem 6.7.1), $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$ if and only if $\mathbb{E}(\exp(\mathrm{i}\boldsymbol{s}'\boldsymbol{z}_n)) \to \mathbb{E}(\exp(\mathrm{i}\boldsymbol{s}'\boldsymbol{z}))$ for every $\boldsymbol{s} \in \mathbb{R}^k$. We can write $\boldsymbol{s} = t\boldsymbol{\lambda}$ where $t \in \mathbb{R}$ and $\boldsymbol{\lambda} \in \mathbb{R}^k$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$. Thus the convergence holds if and only if $\mathbb{E}\left(\exp\left(\mathrm{i}t\boldsymbol{\lambda}'\boldsymbol{z}_n\right)\right) \to \mathbb{E}\left(\exp\left(\mathrm{i}t\boldsymbol{\lambda}'\boldsymbol{z}\right)\right)$ for every $t \in \mathbb{R}$ and $\boldsymbol{\lambda} \in \mathbb{R}^k$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$. Again by Lévy's Continuity Theorem, this holds if and only if $\boldsymbol{\lambda}'\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{\lambda}'\boldsymbol{z}$ for every $\boldsymbol{\lambda} \in \mathbb{R}^k$ and with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$. ■

Proof of Theorem 6.8.1: The moment bound $\mathbb{E}\left(y_i^2\right) < \infty$ is sufficient to guarantee that μ and σ^2 are well defined and finite. Without loss of generality, it is sufficient to consider the case $\mu = 0$.

Our proof method is to calculate the characteristic function of $\sqrt{n}\overline{y}_n$ and show that it converges pointwise to $\exp\left(-\lambda^2\sigma^2/2\right)$, the characteristic function of $\mathrm{N}\left(0, \sigma^2\right)$. By Lévy's Continuity Theorem (Theorem 6.7.1) this implies $\sqrt{n}\overline{y}_n \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\right)$.

Let $C(t) = \mathbb{E}\exp(\mathrm{i}ty_i)$ denote the characteristic function of y_i and set $c(t) = \log C(t)$, which is sometimes called the cumulant generating function. We start by calculating a second order Taylor series expansion of $c(t)$ about $t = 0$, which requires computing the first two derivatives of $c(t)$ at $t = 0$. These derivatives are

$$c'(t) = \frac{C'(t)}{C(t)}$$
$$c''(t) = \frac{C''(t)}{C(t)} - \left(\frac{C'(t)}{C(t)}\right)^2.$$

Using (2.61) and $\mu = 0$ we find

$$c(0) = 0$$
$$c'(0) = 0$$
$$c''(0) = -\sigma^2.$$

Then the second-order Taylor series expansion of $c(t)$ about $t = 0$ equals

$$\begin{aligned} c(t) &= c(0) + c'(0)t + \frac{1}{2}c''(t^*)\boldsymbol{t}^2 \\ &= \frac{1}{2}c''(t^*)t^2 \end{aligned} \tag{6.33}$$

where t^* lies on the line segment joining 0 and t.

We now compute $C_n(t) = \mathbb{E}\exp(\mathrm{i}t\sqrt{n}\overline{y}_n)$, the characteristic function of $\sqrt{n}\overline{y}_n$. By the properties

of the exponential function, the independence of the y_i, and the definition of $c(t)$

$$\begin{aligned}
\log C_n(t) &= \log \mathbb{E}\left(\exp\left(\mathrm{i}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} t y_i\right)\right) \\
&= \log \mathbb{E}\left(\prod_{i=1}^{n} \exp\left(\mathrm{i}\frac{1}{\sqrt{n}} t y_i\right)\right) \\
&= \log \prod_{i=1}^{n} \mathbb{E}\left(\exp\left(\mathrm{i}\frac{1}{\sqrt{n}} t y_i\right)\right) \\
&= \sum_{i=1}^{n} \log \mathbb{E}\left(\exp\left(\mathrm{i}\frac{1}{\sqrt{n}} t y_i\right)\right) \\
&= nc\left(\frac{t}{\sqrt{n}}\right). \\
&= \frac{1}{2} c''(t_n) t^2
\end{aligned}$$

For n large the argument $t/\sqrt{n}$ is in a neighborhood of 0. Since the second moment of y_i is finite, $c''(t)$ is continuous at $t = 0$. Thus we can apply a second order Taylor series expansion about 0, and apply $c(0) = c'(0) = 0$ to find that

$$\begin{aligned}
\log C_n(t) &= nc\left(\frac{t}{\sqrt{n}}\right) \\
&= n\left(c(0) + c'(0)\frac{t}{\sqrt{n}} + \frac{1}{2} c''\left(\frac{t_n}{\sqrt{n}}\right)\left(\frac{t}{\sqrt{n}}\right)^2\right) \\
&= \frac{1}{2} c''\left(\frac{t_n}{\sqrt{n}}\right) t^2
\end{aligned}$$

where t_n lies on the line segment joining 0 and t. Since t_n is bounded we deduce that $c''(t_n/\sqrt{n}) \to c''(0) = -\sigma^2$. Hence, as $n \to \infty$,

$$\log C_n(t) \to -\frac{1}{2}\sigma^2 t^2$$

and

$$C_n(t) \to \exp\left(-\frac{1}{2}\sigma^2 t^2\right)$$

which is the characteristic function of the $\mathrm{N}\left(0, \sigma^2\right)$ distribution, as shown in Exercise 5.9. This completes the proof. ■

Proof of Theorem 6.8.3: Suppose that $\sigma^2 = 0$. Then $\operatorname{var}\left(\sqrt{n}\left(\overline{y} - \mathbb{E}\left(\overline{y}\right)\right)\right) = \overline{\sigma}_n^2 \to \sigma^2 = 0$ so $\sqrt{n}\left(\overline{y} - \mathbb{E}\left(\overline{y}\right)\right) \xrightarrow{p} 0$ and hence $\sqrt{n}\left(\overline{y} - \mathbb{E}\left(\overline{y}\right)\right) \xrightarrow{d} 0$. The random variable $\mathrm{N}\left(0, \sigma^2\right) = \mathrm{N}(0, 0)$ is 0 with probability 1, so this is $\sqrt{n}\left(\overline{y} - \mathbb{E}\left(\overline{y}\right)\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\right)$ as stated.

Now suppose that $\sigma^2 > 0$. This implies (6.8). Together with (6.7) this implies Lyapunov's condition, and hence Lindeberg's condition, and hence Theorem 6.8.2, which states

$$\frac{\sqrt{n}\left(\overline{y} - \mathbb{E}\left(\overline{y}\right)\right)}{\overline{\sigma}_n^{1/2}} \xrightarrow{d} \mathrm{N}(0, 1).$$

Combined with (6.9) we deduce $\sqrt{n}\left(\overline{y} - \mathbb{E}\left(\overline{y}\right)\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2\right)$ as stated. ■

Proof of Theorem 6.9.1: Set $\boldsymbol{\lambda} \in \mathbb{R}^k$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ and define $u_i = \boldsymbol{\lambda}'(\boldsymbol{y}_i - \boldsymbol{\mu})$. The u_i are i.i.d with $\mathbb{E}\left(u_i^2\right) = \boldsymbol{\lambda}'\boldsymbol{V}\boldsymbol{\lambda} < \infty$. By Theorem 6.8.1,

$$\boldsymbol{\lambda}'\sqrt{n}\left(\overline{\boldsymbol{y}} - \boldsymbol{\mu}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i \xrightarrow{d} \mathrm{N}\left(0, \boldsymbol{\lambda}'\boldsymbol{V}\boldsymbol{\lambda}\right)$$

Notice that if $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{V})$ then $\boldsymbol{\lambda}'\boldsymbol{z} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\lambda}'\boldsymbol{V}\boldsymbol{\lambda}\right)$. Thus

$$\boldsymbol{\lambda}'\sqrt{n}\left(\overline{\boldsymbol{y}} - \boldsymbol{\mu}\right) \xrightarrow{d} \boldsymbol{\lambda}'\boldsymbol{z}.$$

Since this holds for all $\boldsymbol{\lambda}$, the conditions of Theorem 6.7.2 are satisfied and we deduce that

$$\sqrt{n}\left(\overline{\boldsymbol{y}} - \boldsymbol{\mu}\right) \xrightarrow{d} \boldsymbol{z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{V})$$

as stated. ■

Proof of Theorem 6.9.2: Set $\boldsymbol{\lambda} \in \mathbb{R}^k$ with with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ and define $u_{ni} = \boldsymbol{\lambda}'\overline{\boldsymbol{V}}_n^{-1/2}\left(\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\right)$. Notice that u_{ni} are independent and has variance $\sigma_{ni}^2 = \boldsymbol{\lambda}'\overline{\boldsymbol{V}}_n^{-1/2}\boldsymbol{V}_{ni}\overline{\boldsymbol{V}}_n^{-1/2}\boldsymbol{\lambda}$ and $\overline{\sigma}_n^2 = n^{-1}\sum_{i=1}^n \sigma_{ni}^2 = 1$. It is sufficient to verify (6.5). By the Cauchy-Schwarz inequality,

$$\begin{aligned} u_{ni}^2 &= \left(\boldsymbol{\lambda}'\overline{\boldsymbol{V}}_n^{-1/2}\left(\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\right)\right)^2 \\ &\leq \boldsymbol{\lambda}'\overline{\boldsymbol{V}}_n^{-1}\boldsymbol{\lambda}\left\|\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\right\|^2 \\ &\leq \frac{\left\|\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\right\|^2}{\lambda_{\min}\left(\overline{\boldsymbol{V}}_n\right)} \\ &= \frac{\left\|\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\right\|^2}{\nu_n^2}. \end{aligned}$$

Then

$$\begin{aligned} \frac{1}{n\overline{\sigma}_n^2}\sum_{i=1}^{n}\mathbb{E}\left(u_{ni}^2 1\left(u_{ni}^2 \geq \varepsilon n\overline{\sigma}_n^2\right)\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(u_{ni}^2 1\left(u_{ni}^2 \geq \varepsilon n\right)\right) \\ &\leq \frac{1}{n\nu_n^2}\sum_{i=1}^{n}\mathbb{E}\left(\left\|\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\right\|^2 1\left(\left\|\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\right\|^2 \geq \varepsilon n\nu_n^2\right)\right) \\ &\to 0 \end{aligned}$$

by (6.11). This establishes (6.5). We deduce from Theorem 6.8.2 that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_{ni} = \boldsymbol{\lambda}'\sqrt{n}\,\overline{\boldsymbol{V}}_n^{-1/2}\left(\overline{\boldsymbol{y}} - \mathbb{E}\left(\overline{\boldsymbol{y}}\right)\right) \xrightarrow{d} \mathrm{N}(0,1) = \boldsymbol{\lambda}'\boldsymbol{z}$$

where $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_k)$. Since this holds for all $\boldsymbol{\lambda}$, the conditions of Theorem 6.7.2 are satisfied and we deduce that

$$\sqrt{n}\,\overline{\boldsymbol{V}}_n^{-1/2}\left(\overline{\boldsymbol{y}} - \mathbb{E}\left(\overline{\boldsymbol{y}}\right)\right) \xrightarrow{d} \mathrm{N}(0, \boldsymbol{I}_k)$$

as stated. ■

Proof of Theorem 6.9.3: Set $\boldsymbol{\lambda} \in \mathbb{R}^k$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ and define $u_{ni} = \boldsymbol{\lambda}'\left(\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\right)$. Using the triangle inequality and (6.13) we obtain

$$\sup_{n,i}\mathbb{E}\left(\left|u_{ni}\right|^{2+\delta}\right) \leq \sup_{n,i}\mathbb{E}\left(\left\|\boldsymbol{y}_{ni} - \boldsymbol{\mu}_{ni}\right\|^{2+\delta}\right) < \infty$$

which is (6.7). Notice that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(u_{ni}^{2}\right)=\boldsymbol{\lambda}'\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{V}_{ni}\boldsymbol{\lambda}=\boldsymbol{\lambda}'\overline{\boldsymbol{V}}_{n}\boldsymbol{\lambda}\rightarrow\boldsymbol{\lambda}'\boldsymbol{V}\boldsymbol{\lambda}$$

which is (6.9). Since the u_{ni} are independent, by Theorem 6.9.1,

$$\boldsymbol{\lambda}'\sqrt{n}\left(\overline{\boldsymbol{y}}-\mathbb{E}\left(\overline{\boldsymbol{y}}\right)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}u_{ni}\xrightarrow{d}\mathrm{N}\left(0,\boldsymbol{\lambda}'\boldsymbol{V}\boldsymbol{\lambda}\right)=\boldsymbol{\lambda}'\boldsymbol{z}$$

where $\boldsymbol{z}\sim\mathrm{N}\left(\boldsymbol{0},\boldsymbol{V}\right)$. Since this holds for all $\boldsymbol{\lambda}$, the conditions of Theorem 6.7.2 are satisfied and we deduce that

$$\sqrt{n}\left(\overline{\boldsymbol{y}}-\mathbb{E}\left(\overline{\boldsymbol{y}}\right)\right)\xrightarrow{d}\mathrm{N}\left(\boldsymbol{0},\boldsymbol{V}\right)$$

as stated. ■

Proof of Theorem 6.12.3: By a vector Taylor series expansion, for each element of $\boldsymbol{g}$,

$$g_{j}(\boldsymbol{\theta}_{n})=g_{j}(\boldsymbol{\theta})+g_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_{jn}^{*})\left(\boldsymbol{\theta}_{n}-\boldsymbol{\theta}\right)$$

where $\boldsymbol{\theta}_{nj}^{*}$ lies on the line segment between $\boldsymbol{\theta}_{n}$ and $\boldsymbol{\theta}$ and therefore converges in probability to $\boldsymbol{\theta}$. It follows that $a_{jn}=g_{j\boldsymbol{\theta}}(\boldsymbol{\theta}_{jn}^{*})-g_{j\boldsymbol{\theta}}\xrightarrow{p}0$. Stacking across elements of $\boldsymbol{g}$, we find

$$\sqrt{n}\left(\boldsymbol{g}\left(\boldsymbol{\theta}_{n}\right)-\boldsymbol{g}(\boldsymbol{\theta})\right)=\left(\boldsymbol{G}+a_{n}\right)'\sqrt{n}\left(\boldsymbol{\theta}_{n}-\boldsymbol{\theta}\right)\xrightarrow{d}\boldsymbol{G}'\boldsymbol{\xi}. \tag{6.34}$$

The convergence is by Theorem 6.12.1, as $\boldsymbol{G}+a_{n}\xrightarrow{d}\boldsymbol{G}$, $\sqrt{n}\left(\boldsymbol{\theta}_{n}-\boldsymbol{\theta}\right)\xrightarrow{d}\boldsymbol{\xi}$, and their product is continuous. This establishes (6.17)

When $\boldsymbol{\xi}\sim\mathrm{N}\left(0,\boldsymbol{V}\right)$, the right-hand-side of (6.34) equals

$$\boldsymbol{G}'\xi=\boldsymbol{G}'\,\mathrm{N}\left(\boldsymbol{0},\boldsymbol{V}\right)=\mathrm{N}\left(\boldsymbol{0},\boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}\right)$$

establishing (6.18). ■

Proof of Theorem 6.14.1: First consider (6.19). Take any $\delta>0$. The event $\left\{\max_{1\leq i\leq n}\left|y_{i}\right|>\delta n^{1/r}\right\}$ means that at least one of the $|y_{i}|$ exceeds $\delta n^{1/r}$, which is the same as the event $\bigcup_{i=1}^{n}\left\{|y_{i}|>\delta n^{1/r}\right\}$ or equivalently $\bigcup_{i=1}^{n}\left\{|y_{i}|^{r}>\delta^{r}n\right\}$. Since the probability of the union of events is smaller than the sum of the probabilities,

$$\begin{aligned}\Pr\left(n^{-1/r}\max_{1\leq i\leq n}\left|y_{i}\right|>\delta\right)&=\Pr\left(\bigcup_{i=1}^{n}\left\{|y_{i}|^{r}>\delta^{r}n\right\}\right)\\&\leq\sum_{i=1}^{n}\Pr\left(|y_{i}|^{r}>n\delta^{r}\right)\\&\leq\frac{1}{n\delta^{r}}\sum_{i=1}^{n}\mathbb{E}\left(|y_{i}|^{r}1\left(|y_{i}|^{r}>n\delta^{r}\right)\right)\\&=\frac{1}{\delta^{r}}\mathbb{E}\left(|y_{i}|^{r}1\left(|y_{i}|^{r}>n\delta^{r}\right)\right)\end{aligned}$$

where the second inequality is the strong form of Markov's inequality (Theorem B.15) and the final equality is since the y_{i} are iid. Since $\mathbb{E}\left(|y|^{r}\right)<\infty$ this final expectation converges to zero as $n\rightarrow\infty$. This is because

$$\mathbb{E}\left(|y_{i}|^{r}\right)=\int|y|^{r}\,dF(y)<\infty$$

implies

$$\mathbb{E}\left(|y_i|^r 1\left(|y_i|^r > c\right)\right) = \int_{|y|^r > c} |y|^r \, dF(y) \to 0 \tag{6.35}$$

as $c \to \infty$. This establishes (6.19).

Now consider (6.20). Take any $\delta > 0$ and pick n large enough so that $(\log n)^\eta t\delta \geq 1$. By a similar calculation

$$\begin{aligned}
\Pr\left((\log n)^{-(1+\eta)} \max_{1 \leq i \leq n} |y_i| > \delta\right) &= \Pr\left(\bigcup_{i=1}^{n} \left\{\exp|ty_i| > \exp\left((\log n)^{1+\eta} t\delta\right)\right\}\right) \\
&\leq \sum_{i=1}^{n} \Pr\left(\exp|ty_i| > n\right) \\
&\leq \mathbb{E}\left(\exp|ty| \, 1\left(\exp|ty| > n\right)\right)
\end{aligned}$$

where the second line uses $\exp\left((\log n)^{1+\eta} t\delta\right) \geq \exp(\log n) = n$. The assumption $\mathbb{E}(\exp(ty)) < \infty$ means $\mathbb{E}\left(\exp|ty| \, 1\left(\exp|ty| > n\right)\right) \to 0$ as $n \to \infty$ by the same argument as in (6.35). This establishes (6.20). ■

Exercises

Exercise 6.1 For the following sequences, show $a_n \to 0$ as $n \to \infty$

(a) $a_n = 1/n$

(b) $a_n = \dfrac{1}{n} \sin\left(\dfrac{\pi}{2} n\right)$

Exercise 6.2 Does the sequence $a_n = \sin\left(\dfrac{\pi}{2} n\right)$ converge? Find the liminf and limsup as $n \to \infty$.

Exercise 6.3 A weighted sample mean takes the form $\overline{y}^* = \frac{1}{n}\sum_{i=1}^n w_i y_i$ for some non-negative constants w_i satisfying $\frac{1}{n}\sum_{i=1}^n w_i = 1$. Assume y_i is iid.

(a) Show that $\overline{y}^*$ is unbiased for $\mu = \mathbb{E}(y_i)$.

(b) Calculate $\text{var}(\overline{y}^*)$.

(c) Show that a sufficient condition for $\overline{y}^* \xrightarrow{p} \mu$ is that $\frac{1}{n^2}\sum_{i=1}^n w_i^2 \longrightarrow 0$.

(d) Show that a sufficient condition for the condition in part 3 is $\max_{i \le n} w_i = o(n)$.

Exercise 6.4 Consider a random variable X_n with the probability distribution

$$X_n = \begin{cases} -n & \text{with probability } 1/n \\ 0 & \text{with probability } 1 - 2/n \\ n & \text{with probability } 1/n \end{cases}$$

(a) Does $X_n \to_p 0$ as $n \to \infty$?

(b) Calculate $\mathbb{E}(X_n)$

(c) Calculate $\text{var}(X_n)$

(d) Now suppose the distribution is

$$X_n = \begin{cases} 0 & \text{with probability } 1 - n \\ n & \text{with probability } 1/n \end{cases}$$

Calculate $\mathbb{E}(X_n)$

(e) Conclude that $X_n \to_p 0$ as $n \to \infty$ and $\mathbb{E}(X_n) \to 0$ are unrelated.

Exercise 6.5 A weighted sample mean takes the form $\overline{y}^* = \frac{1}{n}\sum_{i=1}^n w_i y_i$ for some non-negative constants w_i satisfying $\frac{1}{n}\sum_{i=1}^n w_i = 1$. Assume y_i is iid.

(a) Show that $\overline{y}^*$ is unbiased for $\mu = \mathbb{E}(y_i)$.

(b) Calculate $\text{var}(\overline{y}^*)$.

(c) Show that a sufficient condition for $\overline{y}^* \xrightarrow{p} \mu$ is that $\frac{1}{n^2}\sum_{i=1}^n w_i^2 \longrightarrow 0$.

(d) Show that a sufficient condition for the condition in part c is $\max_{i \le n} w_i / n \to 0$.

Exercise 6.6 Take a random sample $\{y_1, ..., y_n\}$. Which statistics converge in probability by the weak law of large numbers and continuous mapping theorem, assuming the moment exists?

(a) $\frac{1}{n}\sum_{i=1}^n y_i^2$

(b) $\frac{1}{n}\sum_{i=1}^{n} y_i^3$

(c) $\max_{i\leq n} y_i$

(d) $\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2$

(e) $\frac{\sum_{i=1}^{n} y_i^2}{\sum_{i=1}^{n} y_i}$ assuming $\mathbb{E}(y_i) > 0$

(f) $1\left(\frac{1}{n}\sum_{i=1}^{n} y_i > 0\right)$ where

$$1(a) = \begin{cases} 1 & \text{if } a \text{ is true} \\ 0 & \text{if } a \text{ is not true} \end{cases}$$

Exercise 6.7 Take a random sample $\{X_1, ..., X_n\}$ where $X > 0$. Consider the sample geometric mean

$$\widehat{\mu} = \left(\prod_{i=1}^{n} X_i\right)^{1/n}$$

and population geometric mean

$$\mu = \exp\left(\mathbb{E}\left(\log X\right)\right)$$

Assuming μ is finite, show that $\widehat{\mu} \to_p \mu$ as $n \to \infty$.

Exercise 6.8 Take a random variable Z such that $\mathbb{E}(Z) = 0$ and $\text{var}(Z) = 1$. Use Chebyshev's inequality to find a δ such that $\Pr(|Z| > \delta) \leq 0.05$. Contrast this with the exact δ which solves $\Pr(|Z| > \delta) = 0.05$ when $Z \sim \text{N}(0,1)$. Comment on the difference.

Exercise 6.9 Find the moment estimator $\widehat{\mu}_3$ of $\mu_3 = \mathbb{E}\left(y_i^3\right)$ and show that $\sqrt{n}\left(\widehat{\mu}_3 - \mu_3\right) \xrightarrow{d} \text{N}\left(0, v^2\right)$ for some v^2. Write v^2 as a function of the moments of y_i.

Exercise 6.10 Suppose $z_n \xrightarrow{p} c$ as $n \to \infty$. Show that $z_n^2 \xrightarrow{p} c^2$ as $n \to \infty$ using the definition of convergence in probability, but not appealing to the CMT.

Exercise 6.11 Let $\mu_k = \mathbb{E}\left(y^k\right)$ for some integer $k \geq 1$.

(a) Write down the natural moment estimator $\widehat{\mu}_k$ of μ_k.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\mu}_k - \mu_k\right)$ as $n \to \infty$. (Assume $\mathbb{E}\left(X^{2k}\right) < \infty$.)

Exercise 6.12 Let $m_k = \left(\mathbb{E}\left(y^k\right)\right)^{1/k}$ for some integer $k \geq 1$.

(a) Write down an estimator $\widehat{m}_k$ of m_k.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{m}_k - m_k\right)$ as $n \to \infty$.

Exercise 6.13 Suppose $\sqrt{n}\left(\widehat{\mu} - \mu\right) \xrightarrow{d} \text{N}\left(0, v^2\right)$ and set $\beta = \mu^2$ and $\widehat{\beta} = \widehat{\mu}^2$.

(a) Use the Delta Method to obtain an asymptotic distribution for $\sqrt{n}\left(\widehat{\beta} - \beta\right)$.

(b) Now suppose $\mu = 0$. Describe what happens to the asymptotic distribution from the previous part.

(c) Improve on the previous answer. Under the assumption $\mu = 0$, find the asymptotic distribution for $n\widehat{\beta} = n\widehat{\mu}^2$.

(d) Comment on the differences between the answers in parts 1 and 3.

Exercise 6.14 Let y be distributed Bernoulli $P(y = 1) = p$ and $P(y = 0) = 1 - p$ for some unknown $0 < p < 1$.

(a) Show that $p = \mathbb{E}(y)$

(b) Write down the natural moment estimator $\widehat{p}$ of p.

(c) Find $\text{var}(\widehat{p})$

(d) Find the asymptotic distribution of $\sqrt{n}(\widehat{p} - p)$ as $n \to \infty$.

Chapter 7

Asymptotic Theory for Least Squares

7.1 Introduction

It turns out that the asymptotic theory of least-squares estimation applies equally to the projection model and the linear CEF model, and therefore the results in this chapter will be stated for the broader projection model described in Section 2.18. Recall that the model is

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$$

for $i = 1, ..., n$, where the linear projection $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x_i}\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i y_i\right).$$

Some of the results of this section hold under random sampling (Assumption 1.5.2) and finite second moments (Assumption 2.18.1). We restate this condition here for clarity.

Assumption 7.1.1

1. *The observations* $(y_i, \boldsymbol{x_i})$, $i = 1, ..., n$, *are independent and identically distributed.*
2. $\mathbb{E}\left(y^2\right) < \infty$.
3. $\mathbb{E}\left\|\boldsymbol{x}\right\|^2 < \infty$.
4. $\boldsymbol{Q_{xx}} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ *is positive definite.*

Some of the results will require a strengthening to finite fourth moments.

Assumption 7.1.2 *In addition to Assumption 7.1.1,* $\mathbb{E}\left(y_i^4\right) < \infty$ *and* $\mathbb{E}\left\|\boldsymbol{x}_i\right\|^4 < \infty$.

7.2 Consistency of Least-Squares Estimator

In this section we use the weak law of large numbers (WLLN, Theorem 6.4.2 and Theorem 6.6.2) and continuous mapping theorem (CMT, Theorem 6.11.1) to show that the least-squares estimator $\widehat{\boldsymbol{\beta}}$ is consistent for the projection coefficient $\boldsymbol{\beta}$.

This derivation is based on three key components. First, the OLS estimator can be written as a continuous function of a set of sample moments. Second, the WLLN shows that sample moments converge in probability to population moments. And third, the CMT states that continuous functions preserve convergence in probability. We now explain each step in brief and then in greater detail.

First, observe that the OLS estimator

$$\widehat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \widehat{\boldsymbol{Q}}_{\boldsymbol{xy}}$$

is a function of the sample moments $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$ and $\widehat{\boldsymbol{Q}}_{\boldsymbol{xy}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i$.

Second, by an application of the WLLN these sample moments converge in probability to the population moments. Specifically, the fact that $(y_i, \boldsymbol{x}_i)$ are mutually independent and identically distributed implies that any function of $(y_i, \boldsymbol{x}_i)$ is iid, including $\boldsymbol{x}_i \boldsymbol{x}_i'$ and $\boldsymbol{x}_i y_i$. These variables also have finite expectations under Assumption 7.1.1. Under these conditions, the WLLN (Theorem 6.6.2) implies that as $n \to \infty$,

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i'\right) = \boldsymbol{Q}_{\boldsymbol{xx}} \tag{7.1}$$

and

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xy}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i y_i \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i y_i\right) = \boldsymbol{Q}_{\boldsymbol{xy}}. \tag{7.2}$$

Third, the CMT (Theorem 6.11.1) allows us to combine these equations to show that $\widehat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$. Specifically, as $n \to \infty$,

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \widehat{\boldsymbol{Q}}_{\boldsymbol{xy}} \\ &\xrightarrow{p} \boldsymbol{Q}_{\boldsymbol{xx}}^{-1} \boldsymbol{Q}_{\boldsymbol{xy}} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{7.3}$$

We have shown that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$, as $n \to \infty$. In words, the OLS estimator converges in probability to the projection coefficient vector $\boldsymbol{\beta}$ as the sample size n gets large.

To fully understand the application of the CMT we walk through it in detail. We can write

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{g}\left(\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}, \widehat{\boldsymbol{Q}}_{\boldsymbol{xy}}\right)$$

where $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b}) = \boldsymbol{A}^{-1}\boldsymbol{b}$ is a function of $\boldsymbol{A}$ and $\boldsymbol{b}$. The function $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b})$ is a continuous function of $\boldsymbol{A}$ and $\boldsymbol{b}$ at all values of the arguments such that $\boldsymbol{A}^{-1}$ exists. Assumption 7.1.1 specifies that $\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$ exists and thus $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b})$ is continuous at $\boldsymbol{A} = \boldsymbol{Q}_{\boldsymbol{xx}}$. This justifies the application of the CMT in (7.3).

For a slightly different demonstration of (7.3), recall that (4.7) implies that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \widehat{\boldsymbol{Q}}_{\boldsymbol{xe}} \tag{7.4}$$

where

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xe}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i e_i.$$

The WLLN and (2.27) imply

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xe}} \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i e_i\right) = \mathbf{0}. \tag{7.5}$$

Therefore

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \widehat{\boldsymbol{Q}}_{\boldsymbol{xe}} \\ &\xrightarrow{p} \boldsymbol{Q}_{\boldsymbol{xx}}^{-1} \mathbf{0} \\ &= \mathbf{0} \end{aligned}$$

which is the same as $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

Theorem 7.2.1 ***Consistency of Least-Squares***
Under Assumption 7.1.1, $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} \xrightarrow{p} \boldsymbol{Q}_{\boldsymbol{xx}}$, $\widehat{\boldsymbol{Q}}_{\boldsymbol{xy}} \xrightarrow{p} \boldsymbol{Q}_{\boldsymbol{xy}}$, $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \xrightarrow{p} \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$, $\widehat{\boldsymbol{Q}}_{\boldsymbol{xe}} \xrightarrow{p} \mathbf{0}$, *and* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ *as* $n \to \infty$.

Theorem 7.2.1 states that the OLS estimator $\widehat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$ as n increases, and thus $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$. In the stochastic order notation, Theorem 7.2.1 can be equivalently written as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_p(1). \tag{7.6}$$

To illustrate the effect of sample size on the least-squares estimator consider the least-squares regression

$$\ln(Wage_i) = \beta_1 Education_i + \beta_2 Experience_i + \beta_3 Experience_i^2 + \beta_4 + e_i.$$

We use the sample of 24,344 white men from the March 2009 CPS. Randomly sorting the observations, and sequentially estimating the model by least-squares, starting with the first 5 observations, and continuing until the full sample is used, the sequence of estimates are displayed in Figure 7.1. You can see how the least-squares estimate changes with the sample size, but as the number of observations increases it settles down to the full-sample estimate $\widehat{\boldsymbol{\beta}}_1 = 0.114$.

7.3 Asymptotic Normality

We started this chapter discussing the need for an approximation to the distribution of the OLS estimator $\widehat{\boldsymbol{\beta}}$. In Section 7.2 we showed that $\widehat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$. Consistency is a good first step, but in itself does not describe the distribution of the estimator. In this section we derive an approximation typically called the **asymptotic distribution**.

The derivation starts by writing the estimator as a function of sample moments. One of the moments must be written as a sum of zero-mean random vectors and normalized so that the central limit theorem can be applied. The steps are as follows.

Take equation (7.4) and multiply it by $\sqrt{n}$. This yields the expression

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{x}_i e_i\right). \tag{7.7}$$

This shows that the normalized and centered estimator $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ is a function of the sample average $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'$ and the normalized sample average $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{x}_i e_i$. Furthermore, the latter has mean zero so the central limit theorem (CLT, Theorem 6.8.1) applies.

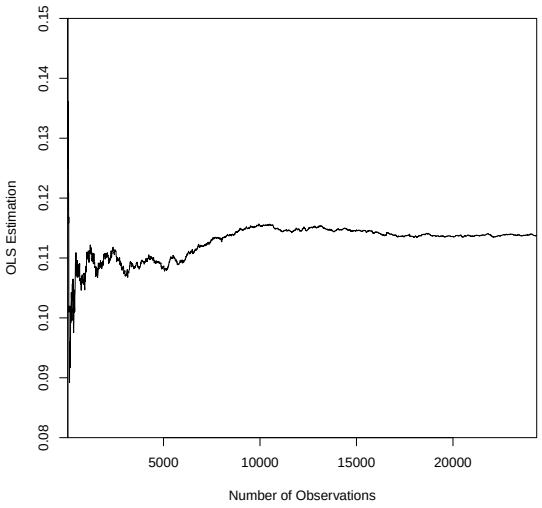


Figure 7.1: The least-squares estimator $\widehat{\boldsymbol{\beta}}_1$ as a function of sample size n

The product $\boldsymbol{x}_i e_i$ is iid (since the observations are iid) and mean zero (since $\mathbb{E}(\boldsymbol{x}_i e_i) = \mathbf{0}$). Define the $k \times k$ covariance matrix

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right). \tag{7.8}$$

We require the elements of $\boldsymbol{\Omega}$ to be finite, written $\boldsymbol{\Omega} < \infty$. It will be useful to recall that Theorem 2.18.1.6 shows that Assumption 7.1.2 implies that $\mathbb{E}\left(e_i^4\right) < \infty$.

The $j\ell^{th}$ element of $\boldsymbol{\Omega}$ is $\mathbb{E}\left(x_{ji} x_{\ell i} e_i^2\right)$. By the Expectation Inequality (B.8), the $j\ell^{th}$ element of $\boldsymbol{\Omega}$ is

$$\left|\mathbb{E}\left(x_{ji} x_{\ell i} e_i^2\right)\right| \le \mathbb{E}\left|x_{ji} x_{\ell i} e_i^2\right| = \mathbb{E}\left(\left|x_{ji}\right| \left|x_{\ell i}\right| e_i^2\right).$$

By two applications of the Cauchy-Schwarz Inequality (B.10), this is smaller than

$$\left(\mathbb{E}\left(x_{ji}^2 x_{\ell i}^2\right)\right)^{1/2} \left(\mathbb{E}\left(e_i^4\right)\right)^{1/2} \le \left(\mathbb{E}\left(x_{ji}^4\right)\right)^{1/4} \left(\mathbb{E}\left(x_{\ell i}^4\right)\right)^{1/4} \left(\mathbb{E}\left(e_i^4\right)\right)^{1/2} < \infty$$

where the finiteness holds under Assumption 7.1.2.

An alternative way to show that the elements of $\boldsymbol{\Omega}$ are finite is by using a matrix norm $\|\cdot\|$ (See Appendix A.18). Then by the Expectation Inequality, the Cauchy-Schwarz Inequality, and Assumption 7.1.2

$$\|\boldsymbol{\Omega}\| \le \mathbb{E}\left\|\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right\| = \mathbb{E}\left(\left\|\boldsymbol{x}_i\right\|^2 e_i^2\right) \le \left(\mathbb{E}\left\|\boldsymbol{x}_i\right\|^4\right)^{1/2} \left(\mathbb{E}\left(e_i^4\right)\right)^{1/2} < \infty.$$

This is a more compact argument (often described as more *elegant*) but such manipulations should not be done without understanding the notation and the applicability of each step of the argument.

Regardless, the finiteness of the covariance matrix means that we can then apply the CLT (Theorem 6.8.1).

Theorem 7.3.1 *Under Assumption 7.1.2,*

$$\boldsymbol{\Omega} < \infty \tag{7.9}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{x}_i e_i \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}) \tag{7.10}$$

as $n \to \infty$.

Putting together (7.1), (7.7), and (7.10),

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &\xrightarrow{d} \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right)\\ &= \mathrm{N}\left(\mathbf{0},\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\right)\end{aligned}$$

as $n\to\infty$, where the final equality follows from the property that linear combinations of normal vectors are also normal (Theorem 5.2.3).

We have derived the asymptotic normal approximation to the distribution of the least-squares estimator.

Theorem 7.3.2 ***Asymptotic Normality of Least-Squares Estimator***
Under Assumption 7.1.2, as $n\to\infty$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}}=\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}, \tag{7.11}$$

$\boldsymbol{Q}_{\boldsymbol{xx}}=\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)$, *and* $\boldsymbol{\Omega}=\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'e_i^2\right)$.

In the stochastic order notation, Theorem 7.3.2 implies that

$$\widehat{\boldsymbol{\beta}}=\boldsymbol{\beta}+O_p(n^{-1/2}) \tag{7.12}$$

which is stronger than (7.6).

The matrix $\boldsymbol{V}_{\boldsymbol{\beta}}=\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$ is the variance of the asymptotic distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$. Consequently, $\boldsymbol{V}_{\boldsymbol{\beta}}$ is often referred to as the **asymptotic covariance matrix** of $\widehat{\boldsymbol{\beta}}$. The expression $\boldsymbol{V}_{\boldsymbol{\beta}}=\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$ is called a **sandwich** form, as the matrix $\boldsymbol{\Omega}$ is sandwiched between two copies of $\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$.

It is useful to compare the variance of the asymptotic distribution given in (7.11) and the finite-sample conditional variance in the CEF model as given in (4.12):

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}=\operatorname{var}\left(\widehat{\boldsymbol{\beta}}\mid\boldsymbol{X}\right)=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{7.13}$$

Notice that $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ is the exact conditional variance of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{V}_{\boldsymbol{\beta}}$ is the asymptotic variance of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$. Thus $\boldsymbol{V}_{\boldsymbol{\beta}}$ should be (roughly) n times as large as $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$, or $\boldsymbol{V}_{\boldsymbol{\beta}}\approx n\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$. Indeed, multiplying (7.13) by n and distributing, we find

$$n\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}=\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

which looks like an estimator of $\boldsymbol{V}_{\boldsymbol{\beta}}$. Indeed, as $n\to\infty$

$$n\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\xrightarrow{p}\boldsymbol{V}_{\boldsymbol{\beta}}.$$

The expression $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ is useful for practical inference (such as computation of standard errors and tests) since it is the variance of the estimator $\widehat{\boldsymbol{\beta}}$, while $\boldsymbol{V}_{\boldsymbol{\beta}}$ is useful for asymptotic theory as it

is well defined in the limit as n goes to infinity. We will make use of both symbols and it will be advisable to adhere to this convention.

There is a special case where $\boldsymbol{\Omega}$ and $\boldsymbol{V}_{\boldsymbol{\beta}}$ simplify. Suppose that

$$\operatorname{cov}(\boldsymbol{x}_i\boldsymbol{x}_i', e_i^2) = \mathbf{0}. \tag{7.14}$$

Condition (7.14) holds in the homoskedastic linear regression model, but is somewhat broader. Under (7.14) the asymptotic variance formulae simplify as

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\mathbb{E}\left(e_i^2\right) = \boldsymbol{Q}_{\boldsymbol{xx}}\sigma^2 \tag{7.15}$$

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1} = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\sigma^2 \equiv \boldsymbol{V}_{\boldsymbol{\beta}}^0 \tag{7.16}$$

In (7.16) we define $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{xx}^{-1}\sigma^2$ whether (7.14) is true or false. When (7.14) is true then $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\beta}}^0$, otherwise $\boldsymbol{V}_{\boldsymbol{\beta}} \neq \boldsymbol{V}_{\boldsymbol{\beta}}^0$. We call $\boldsymbol{V}_{\boldsymbol{\beta}}^0$ the **homoskedastic asymptotic covariance matrix**.

Theorem 7.3.2 states that the sampling distribution of the least-squares estimator, after rescaling, is approximately normal when the sample size n is sufficiently large. This holds true for all joint distributions of $(y_i, \boldsymbol{x}_i)$ which satisfy the conditions of Assumption 7.1.2, and is therefore broadly applicable. Consequently, asymptotic normality is routinely used to approximate the finite sample distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$.

A difficulty is that for any fixed n the sampling distribution of $\widehat{\boldsymbol{\beta}}$ can be arbitrarily far from the normal distribution. In Figure 6.1 we have already seen a simple example where the least-squares estimate is quite asymmetric and non-normal even for reasonably large sample sizes. The normal approximation improves as n increases, but how large should n be in order for the approximation to be useful? Unfortunately, there is no simple answer to this reasonable question. The trouble is that no matter how large is the sample size, the normal approximation is arbitrarily poor for some data distribution satisfying the assumptions. We illustrate this problem using a simulation. Let $y_i = \beta_1 x_i + \beta_2 + e_i$ where x_i is $\mathrm{N}(0,1)$, and e_i is independent of x_i with the Double Pareto density $f(e) = \frac{\alpha}{2}|e|^{-\alpha-1}$, $|e| \geq 1$. If $\alpha > 2$ the error e_i has zero mean and variance $\alpha/(\alpha-2)$. As α approaches 2, however, its variance diverges to infinity. In this context the normalized least-squares slope estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\widehat{\beta}_1 - \beta_1\right)$ has the $\mathrm{N}(0,1)$ asymptotic distribution for any $\alpha > 2$. In Figure 7.2 we display the finite sample densities of the normalized estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\widehat{\beta}_1 - \beta_1\right)$, setting $n = 100$ and varying the parameter α. For $\alpha = 3.0$ the density is very close to the $\mathrm{N}(0,1)$ density. As α diminishes the density changes significantly, concentrating most of the probability mass around zero.

Another example is shown in Figure 7.3. Here the model is $y_i = \beta + e_i$ where

$$e_i = \frac{u_i^r - \mathbb{E}\left(u_i^r\right)}{\left(\mathbb{E}\left(u_i^{2r}\right) - \left(\mathbb{E}\left(u_i^r\right)\right)^2\right)^{1/2}} \tag{7.17}$$

and $u_i \sim \mathrm{N}(0,1)$ and some integer $r \geq 1$. We show the sampling distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ setting $n = 100$, for $r = 1$, 4, 6 and 8. As r increases, the sampling distribution becomes highly skewed and non-normal. The lesson from Figures 7.2 and 7.3 is that the $\mathrm{N}(0,1)$ asymptotic approximation is never guaranteed to be accurate.

7.4 Joint Distribution

Theorem 7.3.2 gives the joint asymptotic distribution of the coefficient estimates. We can use the result to study the covariance between the coefficient estimates. For simplicity, suppose $k = 2$ with no intercept, both regressors are mean zero and the error is homoskedastic. Let σ_1^2 and σ_2^2 be

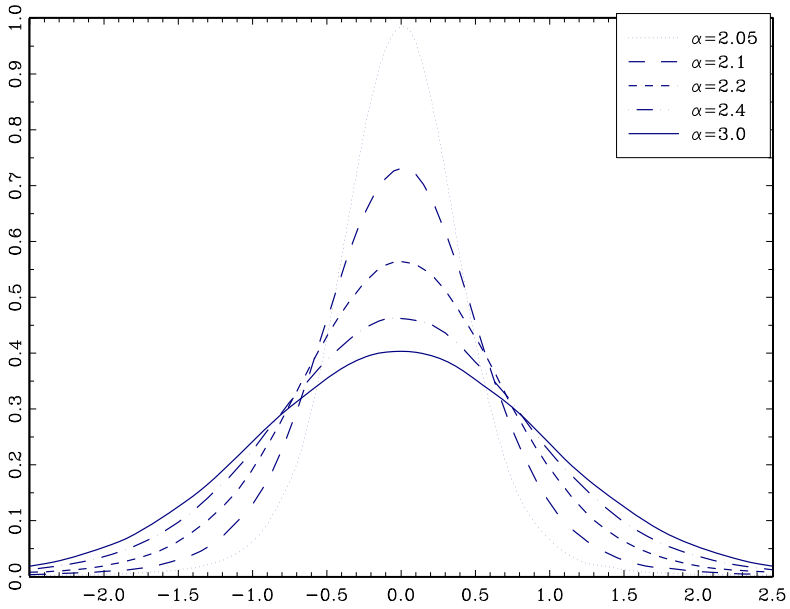


Figure 7.2: Density of Normalized OLS estimator with Double Pareto Error

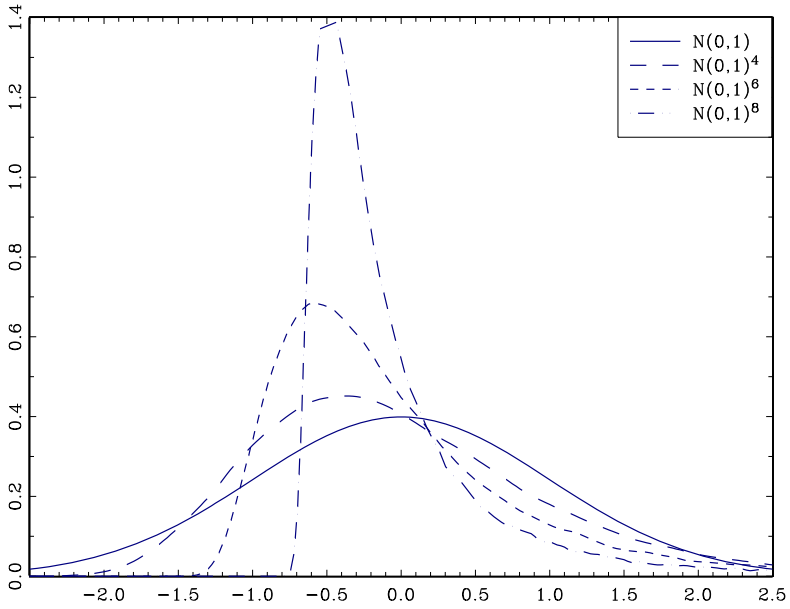


Figure 7.3: Density of Normalized OLS estimator with error process (7.17)

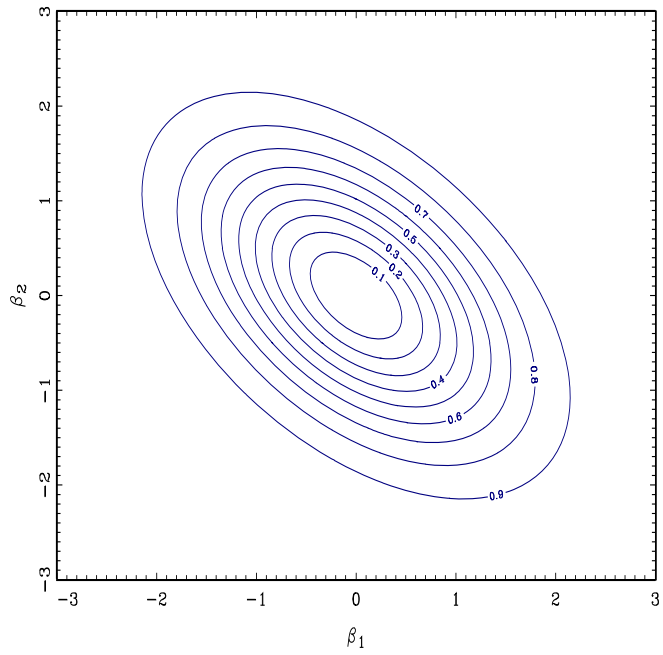


Figure 7.4: Contours of Joint Distribution of $(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$, homoskedastic case

the variances of x_{1i} and x_{2i}, and ρ be their correlation. Then using the formula for inversion of a 2×2 matrix,

$$\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \sigma^2 \boldsymbol{Q}_{\boldsymbol{xx}}^{-1} = \frac{\sigma^2}{\sigma_1^2 \sigma_2^2 \left(1-\rho^2\right)} \begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix}.$$

Thus if x_{1i} and x_{2i} are positively correlated ($\rho > 0$) then $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are negatively correlated (and vice-versa).

For illustration, Figure 7.4 displays the probability contours of the joint asymptotic distribution of $\widehat{\beta}_1 - \beta_1$ and $\widehat{\beta}_2 - \beta_2$ when $\beta_1 = \beta_2 = 0$, $\sigma_1^2 = \sigma_2^2 = \sigma^2 = 1$, and $\rho = 0.5$. The coefficient estimates are negatively correlated since the regressors are positively correlated. This means that if $\widehat{\beta}_1$ is unusually negative, it is likely that $\widehat{\beta}_2$ is unusually positive, or conversely. It is also unlikely that we will observe both $\widehat{\beta}_1$ and $\widehat{\beta}_2$ unusually large and of the same sign.

This finding that the correlation of the regressors is of opposite sign of the correlation of the coefficient estimates is sensitive to the assumption of homoskedasticity. If the errors are heteroskedastic then this relationship is not guaranteed.

This can be seen through a simple constructed example. Suppose that x_{1i} and x_{2i} only take the values $\{-1, +1\}$, symmetrically, with $\Pr(x_{1i} = x_{2i} = 1) = \Pr(x_{1i} = x_{2i} = -1) = 3/8$, and $\Pr(x_{1i} = 1, x_{2i} = -1) = \Pr(x_{1i} = -1, x_{2i} = 1) = 1/8$. You can check that the regressors are mean zero, unit variance and correlation 0.5, which is identical with the setting displayed in Figure 7.4.

Now suppose that the error is heteroskedastic. Specifically, suppose that $\mathbb{E}\left(e_i^2 \mid x_{1i} = x_{2i}\right) = \dfrac{5}{4}$ and $\mathbb{E}\left(e_i^2 \mid x_{1i} \neq x_{2i}\right) = \dfrac{1}{4}$. You can check that $\mathbb{E}\left(e_i^2\right) = 1$, $\mathbb{E}\left(x_{1i}^2 e_i^2\right) = \mathbb{E}\left(x_{2i}^2 e_i^2\right) = 1$ and

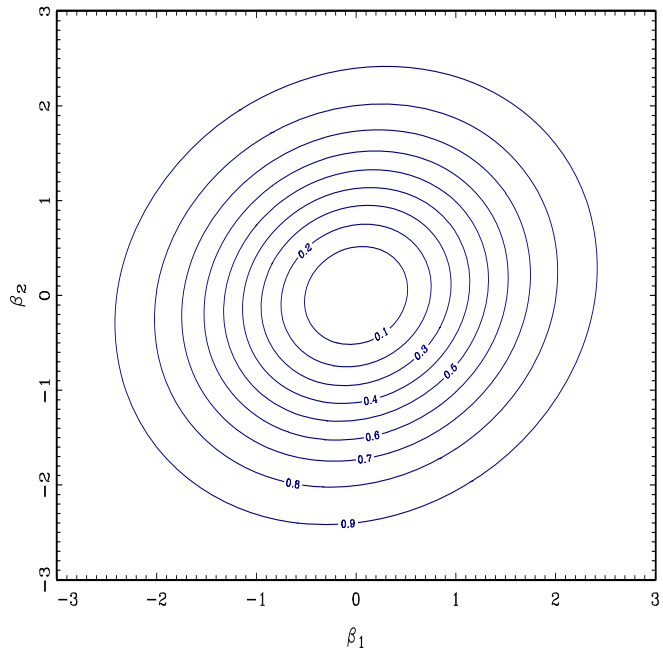


Figure 7.5: Contours of Joint Distribution of $\widehat{\beta}_1$ and $\widehat{\beta}_2$, heteroskedastic case

$\mathbb{E}\left(x_{1i}x_{2i}e_i^2\right) = \dfrac{7}{8}$. Therefore

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{\beta}} &= \boldsymbol{Q}_{xx}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{xx}^{-1} \\ &= \frac{9}{16}\begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}\begin{bmatrix} 1 & \frac{7}{8} \\ \frac{7}{8} & 1 \end{bmatrix}\begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \\ &= \frac{4}{3}\begin{bmatrix} 1 & \frac{1}{4} \\ \frac{1}{4} & 1 \end{bmatrix}.\end{aligned}$$

Thus the coefficient estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are positively correlated (their correlation is 1/4.) The joint probability contours of their asymptotic distribution is displayed in Figure 7.5. We can see how the two estimates are positively associated.

What we found through this example is that in the presence of heteroskedasticity there is no simple relationship between the correlation of the regressors and the correlation of the parameter estimates.

We can extend the above analysis to study the covariance between coefficient sub-vectors. For example, partitioning $\boldsymbol{x}_i' = (\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}')$ and $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)$, we can write the general model as

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i$$

and the coefficient estimates as $\widehat{\boldsymbol{\beta}}' = \left(\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2'\right)$. Make the partitions

$$\boldsymbol{Q}_{\boldsymbol{xx}} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}, \qquad \boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{bmatrix}. \tag{7.18}$$

From (2.41)

$$\boldsymbol{Q}_{\boldsymbol{xx}}^{-1} = \begin{bmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{bmatrix}$$

where $\boldsymbol{Q}_{11\cdot2} = \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and $\boldsymbol{Q}_{22\cdot1} = \boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$. Thus when the error is homoskedastic,

$$\operatorname{cov}\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2\right) = -\sigma^2 \boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}$$

which is a matrix generalization of the two-regressor case.

In the general case, you can show that (Exercise 7.5)

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix} \tag{7.19}$$

where

$$\boldsymbol{V}_{11} = \boldsymbol{Q}_{11\cdot2}^{-1}\left(\boldsymbol{\Omega}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{\Omega}_{21} - \boldsymbol{\Omega}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{\Omega}_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot2}^{-1} \tag{7.20}$$

$$\boldsymbol{V}_{21} = \boldsymbol{Q}_{22\cdot1}^{-1}\left(\boldsymbol{\Omega}_{21} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot2}^{-1} \tag{7.21}$$

$$\boldsymbol{V}_{22} = \boldsymbol{Q}_{22\cdot1}^{-1}\left(\boldsymbol{\Omega}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{12} - \boldsymbol{\Omega}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12} + \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{11}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\right)\boldsymbol{Q}_{22\cdot1}^{-1} \tag{7.22}$$

Unfortunately, these expressions are not easily interpretable.

7.5 Consistency of Error Variance Estimators

Using the methods of Section 7.2 we can show that the estimators $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{e}_i^2$ and $s^2 = \frac{1}{n-k}\sum_{i=1}^n \widehat{e}_i^2$ are consistent for σ^2.

The trick is to write the residual $\widehat{e}_i$ as equal to the error e_i plus a deviation term

$$\begin{aligned} \widehat{e}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \\ &= e_i + \boldsymbol{x}_i'\boldsymbol{\beta} - x_i'\widehat{\boldsymbol{\beta}} \\ &= e_i - \boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \end{aligned}$$

Thus the squared residual equals the squared error plus a deviation

$$\widehat{e}_i^2 = e_i^2 - 2e_i\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{x}_i\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{7.23}$$

So when we take the average of the squared residuals we obtain the average of the squared errors, plus two terms which are (hopefully) asymptotically negligible.

$$\begin{aligned} \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^n e_i\boldsymbol{x}_i'\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \end{aligned} \tag{7.24}$$

Indeed, the WLLN shows that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n e_i^2 &\xrightarrow{p} \sigma^2 \\ \frac{1}{n}\sum_{i=1}^n e_i\boldsymbol{x}_i' &\xrightarrow{p} \mathbb{E}\left(e_i\boldsymbol{x}_i'\right) = 0 \\ \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' &\xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) = \boldsymbol{Q}_{\boldsymbol{xx}} \end{aligned}$$

and Theorem 7.2.1 shows that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$. Hence (7.24) converges in probability to σ^2, as desired.

Finally, since $n/(n-k) \to 1$ as $n \to \infty$, it follows that

$$s^2 = \left(\frac{n}{n-k}\right)\widehat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

Thus both estimators are consistent.

Theorem 7.5.1 *Under Assumption 7.1.1, $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$ and $s^2 \xrightarrow{p} \sigma^2$ as $n \to \infty$.*

7.6 Homoskedastic Covariance Matrix Estimation

Theorem 7.3.2 shows that $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ is asymptotically normal with asymptotic covariance matrix $\boldsymbol{V}_{\boldsymbol{\beta}}$. For asymptotic inference (confidence intervals and tests) we need a consistent estimate of $\boldsymbol{V}_{\boldsymbol{\beta}}$. Under homoskedasticity, $\boldsymbol{V}_{\boldsymbol{\beta}}$ simplifies to $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\sigma^2$, and in this section we consider the simplified problem of estimating $\boldsymbol{V}_{\boldsymbol{\beta}}^0$.

The standard moment estimator of $\boldsymbol{Q}_{\boldsymbol{xx}}$ is $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}$ defined in (7.1), and thus an estimator for $\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$ is $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}$. Also, the standard estimator of σ^2 is the unbiased estimator s^2 defined in (4.30). Thus a natural plug-in estimator for $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\sigma^2$ is $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}s^2$.

Consistency of $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0$ for $\boldsymbol{V}_{\boldsymbol{\beta}}^0$ follows from consistency of the moment estimates $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}$ and s^2, and an application of the continuous mapping theorem. Specifically, Theorem 7.2.1 established that $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} \xrightarrow{p} \boldsymbol{Q}_{\boldsymbol{xx}}$, and Theorem 7.5.1 established $s^2 \xrightarrow{p} \sigma^2$. The function $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\sigma^2$ is a continuous function of $\boldsymbol{Q}_{\boldsymbol{xx}}$ and σ^2 so long as $\boldsymbol{Q}_{\boldsymbol{xx}} > 0$, which holds true under Assumption 7.1.1.4. It follows by the CMT that

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}s^2 \xrightarrow{p} \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\sigma^2 = \boldsymbol{V}_{\boldsymbol{\beta}}^0$$

so that $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0$ is consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}^0$, as desired.

Theorem 7.6.1 *Under Assumption 7.1.1, $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}^0$ as $n \to \infty$.*

It is instructive to notice that Theorem 7.6.1 does not require the assumption of homoskedasticity. That is, $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0$ is consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}^0$ regardless if the regression is homoskedastic or heteroskedastic. However, $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{V}_{\boldsymbol{\beta}} = \text{avar}(\widehat{\boldsymbol{\beta}})$ only under homoskedasticity. Thus in the general case, $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0$ is consistent for a well-defined but non-useful object.

7.7 Heteroskedastic Covariance Matrix Estimation

Theorems 7.3.2 established that the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ is $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$. We now consider estimation of this covariance matrix without imposing homoskedasticity. The standard approach is to use a plug-in estimator which replaces the unknowns with sample moments.

As described in the previous section, a natural estimator for $\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$ is $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}$, where $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}$ defined in (7.1).

The moment estimator for $\boldsymbol{\Omega}$ is

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \widehat{e}_i^2, \tag{7.25}$$

leading to the plug-in covariance matrix estimator

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W} = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \widehat{\boldsymbol{\Omega}} \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}. \tag{7.26}$$

You can check that $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W} = n\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ where $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ is the White covariance matrix estimator introduced in (4.37).

As shown in Theorem 7.2.1, $\widehat{\boldsymbol{Q}}_{xx}^{-1} \xrightarrow{p} \boldsymbol{Q}_{xx}^{-1}$, so we just need to verify the consistency of $\widehat{\boldsymbol{\Omega}}$. The key is to replace the squared residual $\widehat{e}_i^2$ with the squared error e_i^2, and then show that the difference is asymptotically negligible.

Specifically, observe that

$$\begin{aligned} \widehat{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \widehat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left(\widehat{e}_i^2 - e_i^2\right). \end{aligned} \tag{7.27}$$

The first term is an average of the iid random variables $\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2$, and therefore by the WLLN converges in probability to its expectation, namely,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right) = \boldsymbol{\Omega}.$$

Technically, this requires that $\boldsymbol{\Omega}$ has finite elements, which was shown in (7.9).

So to establish that $\widehat{\boldsymbol{\Omega}}$ is consistent for $\boldsymbol{\Omega}$ it remains to show that

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left(\widehat{e}_i^2 - e_i^2\right) \xrightarrow{p} 0. \tag{7.28}$$

There are multiple ways to do this. A reasonable straightforward yet slightly tedious derivation is to start by applying the Triangle Inequality (A.26) using a matrix norm:

$$\begin{aligned} \left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| &\le \frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{x}_i \boldsymbol{x}_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| \\ &= \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 \left|\widehat{e}_i^2 - e_i^2\right|. \end{aligned} \tag{7.29}$$

Then recalling the expression for the squared residual (7.23), apply the Triangle Inequality and then the Schwarz Inequality (A.20) twice

$$\begin{aligned} \left|\widehat{e}_i^2 - e_i^2\right| &\le 2\left|e_i \boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right| + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{x}_i \boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &= 2\left|e_i\right| \left|\boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right| + \left|\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{x}_i\right|^2 \\ &\le 2\left|e_i\right| \|\boldsymbol{x}_i\| \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| + \|\boldsymbol{x}_i\|^2 \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2. \end{aligned} \tag{7.30}$$

Combining (7.29) and (7.30), we find

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\left(\widehat{e}_i^2-e_i^2\right)\right\| &\leq 2\left(\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^3|e_i|\right)\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\| \\ &\quad+\left(\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^4\right)\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|^2 \\ &= o_p(1). \end{aligned} \tag{7.31}$$

The expression is $o_p(1)$ because $\left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\| \xrightarrow{p} 0$ and both averages in parenthesis are averages of random variables with finite mean under Assumption 7.1.2 (and are thus $O_p(1)$). Indeed, by Hölder's Inequality (B.9)

$$\begin{aligned}\mathbb{E}\left(\|\boldsymbol{x}_i\|^3|e_i|\right) &\leq \left(\mathbb{E}\left(\|\boldsymbol{x}_i\|^3\right)^{4/3}\right)^{3/4}\left(\mathbb{E}\left(e_i^4\right)\right)^{1/4} \\ &= \left(\mathbb{E}\left(\|\boldsymbol{x}_i\|^4\right)\right)^{3/4}\left(\mathbb{E}\left(e_i^4\right)\right)^{1/4}<\infty.\end{aligned}$$

We have established (7.28), as desired.

Theorem 7.7.1 *Under Assumption 7.1.2, as* $n\to\infty$, $\widehat{\boldsymbol{\Omega}}\xrightarrow{p}\boldsymbol{\Omega}$ *and* $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W}\xrightarrow{p}\boldsymbol{V}_{\boldsymbol{\beta}}$.

For an alternative proof of this result, see Section 7.21.

7.8 Summary of Covariance Matrix Notation

The notation we have introduced may be somewhat confusing so it is helpful to write it down in one place. The exact variance of $\widehat{\boldsymbol{\beta}}$ (under the assumptions of the linear regression model) and the asymptotic variance of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ (under the more general assumptions of the linear projection model) are

$$\begin{aligned}\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} &= \operatorname{var}\left(\widehat{\boldsymbol{\beta}}\mid\boldsymbol{X}\right)=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ \boldsymbol{V}_{\boldsymbol{\beta}} &= \operatorname{avar}\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right)=\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}.\end{aligned}$$

The White estimates of these two covariance matrices are

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W} &= \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\end{aligned}$$

and satisfy the simple relationship

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W}=n\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}.$$

Similarly, under the assumption of homoskedasticity the exact and asymptotic variances simplify to

$$\begin{aligned}\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}^{0} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\ \boldsymbol{V}_{\boldsymbol{\beta}}^{0} &= \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\sigma^2\end{aligned}$$

and their standard estimators are

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} s^2$$
$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0} = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} s^2$$

which also satisfy the relationship

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0} = n\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}.$$

The exact formula and estimates are useful when constructing test statistics and standard errors. However, for theoretical purposes the asymptotic formula (variances and their estimates) are more useful, as these retain non-generate limits as the sample sizes diverge. That is why both sets of notation are useful.

7.9 Alternative Covariance Matrix Estimators*

In Section 7.7 we introduced $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W}$ as an estimator of $\boldsymbol{V}_{\boldsymbol{\beta}}$. $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W}$ is a scaled version of $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ from Section 4.13, where we also introduced the alternative heteroskedasticity-robust covariance matrix estimators $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ and $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$. We now discuss the consistency properties of these estimators.

To do so we introduce their scaled versions, e.g. $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = n\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}} = n\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, and $\overline{\boldsymbol{V}}_{\boldsymbol{\beta}} = n\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$. These are (alternative) estimates of the asymptotic covariance matrix $\boldsymbol{V}_{\boldsymbol{\beta}}$.

First, consider $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$. Notice that $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = n\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \frac{n}{n-k}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W}$ where $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W}$ was defined in (7.26) and shown consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}$ in Theorem 7.7.1. If k is fixed as $n \to \infty$, then $\frac{n}{n-k} \to 1$ and thus

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = (1 + o(1))\,\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}.$$

Thus $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}$.

The alternative estimators $\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}$ and $\overline{\boldsymbol{V}}_{\boldsymbol{\beta}}$ take the form (7.26) but with $\widehat{\boldsymbol{\Omega}}$ replaced by

$$\widetilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} (1 - h_{ii})^{-2}\, \boldsymbol{x}_i \boldsymbol{x}_i' \widehat{e}_i^2$$

and

$$\overline{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} (1 - h_{ii})^{-1}\, \boldsymbol{x}_i \boldsymbol{x}_i' \widehat{e}_i^2,$$

respectively. To show that these estimators also consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}$, given $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$, it is sufficient to show that the differences $\widetilde{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}}$ and $\overline{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}}$ converge in probability to zero as $n \to \infty$.

The trick is to use the fact that the leverage values are asymptotically negligible:

$$h_n^* = \max_{1 \le i \le n} h_{ii} = o_p(1). \tag{7.32}$$

(See Theorem 7.22.1 in Section 7.22).) Then using the Triangle Inequality

$$\begin{aligned}
\left\| \overline{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}} \right\| &\le \frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{x}_i \boldsymbol{x}_i' \right\| \widehat{e}_i^2 \left| (1 - h_{ii})^{-1} - 1 \right| \\
&\le \left(\frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{x}_i \right\|^2 \widehat{e}_i^2 \right) \left| (1 - h_n^*)^{-1} - 1 \right|.
\end{aligned}$$

The sum in parenthesis can be shown to be $O_p(1)$ under Assumption 7.1.2 by the same argument as in in the proof of Theorem 7.7.1. (In fact, it can be shown to converge in probability to

$\mathbb{E}\left(\|\boldsymbol{x}_i\|^2 e_i^2\right)$.) The term in absolute values is $o_p(1)$ by (7.32). Thus the product is $o_p(1)$, which means that $\overline{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}} + o_p(1) \longrightarrow \boldsymbol{\Omega}$.

Similarly,

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}}\right\| &\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\boldsymbol{x}_i'\right\|\widehat{e}_i^2\left|(1-h_{ii})^{-2}-1\right| \\ &\leq \left(\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{x}_i\|^2\widehat{e}_i^2\right)\left|(1-h_n^*)^{-2}-1\right| \\ &= o_p(1).\end{aligned}$$

Theorem 7.9.1 *Under Assumption 7.1.2, as* $n \to \infty$, $\widetilde{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$, $\overline{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$, $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$, *and* $\overline{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$.

Theorem 7.9.1 shows that the alternative covariance matrix estimators are also consistent for the asymptotic covariance matrix.

7.10 Functions of Parameters

In most serious applications the researcher is actually interested in a specific transformation of the coefficient vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)$. For example, he or she may be interested in a single coefficient β_j, or a ratio β_j/β_l. More generally, interest may focus on a quantity such as consumer surplus which could be a complicated function of the coefficients. In any of these cases we can write the parameter of interest $\boldsymbol{\theta}$ as a function of the coefficients, e.g. $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$ for some function $\boldsymbol{r} : \mathbb{R}^k \to \mathbb{R}^q$. The estimate of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{r}(\widehat{\boldsymbol{\beta}}).$$

By the continuous mapping theorem (Theorem 6.11.1) and the fact $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ we can deduce that $\widehat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$ (if the function $\boldsymbol{r}(\cdot)$ is continuous).

Theorem 7.10.1 *Under Assumption 7.1.1, if* $\boldsymbol{r}(\boldsymbol{\beta})$ *is continuous at the true value of* $\boldsymbol{\beta}$, *then as* $n \to \infty$, $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$.

Furthermore, if the transformation is sufficiently smooth, by the Delta Method (Theorem 6.12.3) we can show that $\widehat{\boldsymbol{\theta}}$ is asymptotically normal.

Assumption 7.10.1 $\boldsymbol{r}(\boldsymbol{\beta}) : \mathbb{R}^k \to \mathbb{R}^q$ *is continuously differentiable at the true value of* $\boldsymbol{\beta}$ *and* $\boldsymbol{R} = \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}(\boldsymbol{\beta})'$ *has rank* q.

Theorem 7.10.2 ***Asymptotic Distribution of Functions of Parameters***
Under Assumptions 7.1.2 and 7.10.1, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right) \tag{7.33}$$

where

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R} \tag{7.34}$$

In many cases, the function $\boldsymbol{r}(\boldsymbol{\beta})$ is linear:

$$\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta}$$

for some $k \times q$ matrix $\boldsymbol{R}$. In particular, if $\boldsymbol{R}$ is a "selector matrix"

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{I} \\ \mathbf{0} \end{pmatrix} \tag{7.35}$$

then we can partition $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ so that $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\beta}_1$ for $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. Then

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \end{pmatrix} \boldsymbol{V}_{\boldsymbol{\beta}} \begin{pmatrix} \boldsymbol{I} \\ \mathbf{0} \end{pmatrix} = \boldsymbol{V}_{11},$$

the upper-left sub-matrix of $\boldsymbol{V}_{11}$ given in (7.20). In this case (7.33) states that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{11}\right).$$

That is, subsets of $\widehat{\boldsymbol{\beta}}$ are approximately normal with variances given by the conformable subcomponents of $\boldsymbol{V}$.

To illustrate the case of a nonlinear transformation, take the example $\theta = \beta_j/\beta_l$ for $j \neq l$. Then

$$\boldsymbol{R} = \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial}{\partial \beta_1}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_j}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_\ell}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_k}\left(\beta_j/\beta_l\right) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1/\beta_l \\ \vdots \\ -\beta_j/\beta_l^2 \\ \vdots \\ 0 \end{pmatrix} \tag{7.36}$$

so

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{V}_{jj}/\beta_l^2 + \boldsymbol{V}_{ll}\beta_j^2/\beta_l^4 - 2\boldsymbol{V}_{jl}\beta_j/\beta_l^3$$

where $\boldsymbol{V}_{ab}$ denotes the ab^{th} element of $\boldsymbol{V}_{\boldsymbol{\beta}}$.

For inference we need an estimate of the asymptotic variance matrix $\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}$, and for this it is typical to use a plug-in estimator. The natural estimator of $\boldsymbol{R}$ is the derivative evaluated at the point estimates

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}(\widehat{\boldsymbol{\beta}})'. \tag{7.37}$$

The derivative in (7.37) may be calculated analytically or numerically. By analytically, we mean working out for the formula for the derivative and replacing the unknowns by point estimates. For

example, if $\theta = \beta_j/\beta_l$, then $\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}(\boldsymbol{\beta})$ is (7.36). However in some cases the function $\boldsymbol{r}(\boldsymbol{\beta})$ may be extremely complicated and a formula for the analytic derivative may not be easily available. In this case calculation by numerical differentiation may be preferable. Let $\delta_l = (0 \cdots 1 \cdots 0)'$ be the unit vector with the "1" in the l^{th} place. Then the jl'th element of a numerical derivative $\widehat{\boldsymbol{R}}$ is

$$\widehat{\boldsymbol{R}}_{jl} = \frac{\boldsymbol{r}_j(\widehat{\boldsymbol{\beta}} + \delta_l \varepsilon) - \boldsymbol{r}_j(\widehat{\boldsymbol{\beta}})}{\varepsilon}$$

for some small ε.

The estimate of $\boldsymbol{V}_{\boldsymbol{\theta}}$ is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}. \tag{7.38}$$

Alternatively, $\widehat{\boldsymbol{V}}^0_{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}$ or $\overline{\boldsymbol{V}}_{\boldsymbol{\beta}}$ may be used in place of $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$. For example, the homoskedastic covariance matrix estimator is

$$\widehat{\boldsymbol{V}}^0_{\boldsymbol{\theta}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}^{\boldsymbol{0}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{Q}}^{-1}_{\boldsymbol{xx}} \widehat{\boldsymbol{R}} s^2 \tag{7.39}$$

Given (7.37), (7.38) and (7.39) are simple to calculate using matrix operations.

As the primary justification for $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ is the asymptotic approximation (7.33), $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ is often called an **asymptotic covariance matrix estimator**.

The estimator $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{V}_{\boldsymbol{\theta}}$ under the conditions of Theorem 7.10.2 since $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$ by Theorem 7.7.1, and

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\widehat{\boldsymbol{\beta}})' \xrightarrow{p} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\boldsymbol{\beta})' = \boldsymbol{R}$$

since $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ and the function $\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}(\boldsymbol{\beta})'$ is continuous in $\boldsymbol{\beta}$.

Theorem 7.10.3 *Under Assumptions 7.1.2 and 7.10.1, as $n \to \infty$,*

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\theta}}.$$

Theorem 7.10.3 shows that $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{V}_{\boldsymbol{\theta}}$ and thus may be used for asymptotic inference. In practice, we may set

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}} = n^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}} \tag{7.40}$$

as an estimate of the variance of $\widehat{\boldsymbol{\theta}}$, or substitute an alternative covariance estimator such as $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$.

7.11 Asymptotic Standard Errors

As described in Section 4.14, a standard error is an estimate of the standard deviation of the distribution of an estimator. Thus if $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is an estimate of the covariance matrix of $\widehat{\boldsymbol{\beta}}$, then standard errors are the square roots of the diagonal elements of this matrix. These take the form

$$s(\widehat{\beta}_j) = \sqrt{\widehat{\boldsymbol{V}}_{\widehat{\beta}_j}} = \sqrt{\left[\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\right]_{jj}}.$$

Standard errors for $\widehat{\boldsymbol{\theta}}$ are constructed similarly. Supposing that $q = 1$ (so $h(\boldsymbol{\beta})$ is real-valued), then the standard error for $\widehat{\theta}$ is the square root of (7.40)

$$s(\widehat{\theta}) = \sqrt{\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}} = \sqrt{n^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}}.$$

When the justification is based on asymptotic theory we call $s(\widehat{\beta}_j)$ or $s(\widehat{\theta})$ an **asymptotic standard error** for $\widehat{\beta}_j$ or $\widehat{\theta}$. When reporting your results, it is good practice to report standard errors for each reported estimate, and this includes functions and transformations of your parameter estimates. This helps users of the work (including yourself) assess the estimation precision.

We illustrate using the log wage regression

$$\log(Wage) = \beta_1\, education + \beta_2\, experience + \beta_3\, experience^2/100 + \beta_4 + e.$$

Consider the following three parameters of interest.

1. Percentage return to education:
$$\theta_1 = 100\beta_1$$
(100 times the partial derivative of the conditional expectation of log wages with respect to *education*.)

2. Percentage return to experience for individuals with 10 years of experience:
$$\theta_2 = 100\beta_2 + 20\beta_3$$
(100 times the partial derivative of the conditional expectation of log wages with respect to *experience*, evaluated at *experience* = 10.)

3. Experience level which maximizes expected log wages:
$$\theta_3 = -50\beta_2/\beta_3$$
(The level of *experience* at which the partial derivative of the conditional expectation of log wages with respect to *experience* equals 0.)

The 4×1 vector $\boldsymbol{R}$ for these three parameters is

$$\boldsymbol{R} = \begin{pmatrix} 100 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 \\ 100 \\ 20 \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} 0 \\ -50/\beta_3 \\ 50\beta_2/\beta_3^2 \\ 0 \end{pmatrix},$$

respectively.

We use the subsample of married black women (all experience levels), which has 982 observations. The point estimates and standard errors are

$$\widehat{\log(Wage)} = \underset{(0.008)}{0.118}\ education + \underset{(0.006)}{0.016}\ experience - \underset{(0.012)}{0.022}\ experience^2/100 + \underset{(0.157)}{0.947}. \tag{7.41}$$

The standard errors are the square roots of the Horn-Horn-Duncan covariance matrix estimate

$$\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \begin{pmatrix} 0.632 & 0.131 & -0.143 & -11.1 \\ 0.131 & 0.390 & -0.731 & -6.25 \\ -0.143 & -0.731 & 1.48 & 9.43 \\ -11.1 & -6.25 & 9.43 & 246 \end{pmatrix} \times 10^{-4}. \tag{7.42}$$

We calculate that

$$\begin{aligned} \widehat{\theta}_1 &= 100\widehat{\beta}_1 \\ &= 100 \times 0.118 \\ &= 11.8 \end{aligned}$$

$$
\begin{aligned}
s(\widehat{\theta}_1) &= \sqrt{100^2 \times 0.632 \times 10^{-4}} \\
&= 0.8
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\theta}_2 &= 100\widehat{\beta}_2 + 20\widehat{\beta}_3 \\
&= 100 \times 0.016 - 20 \times 0.022 \\
&= 1.16
\end{aligned}
$$

$$
\begin{aligned}
s(\widehat{\theta}_2) &= \sqrt{\begin{pmatrix} 100 & 20 \end{pmatrix} \begin{pmatrix} 0.390 & -0.731 \\ -0.731 & 1.48 \end{pmatrix} \begin{pmatrix} 100 \\ 20 \end{pmatrix} \times 10^{-4}} \\
&= 0.55
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\theta}_3 &= -50\widehat{\beta}_2/\widehat{\beta}_3 \\
&= 50 \times 0.016/0.022 \\
&= 35.2
\end{aligned}
$$

$$
\begin{aligned}
s(\widehat{\theta}_3) &= \sqrt{\begin{pmatrix} -50/\widehat{\beta}_3 & 50\widehat{\beta}_2/\widehat{\beta}_3^2 \end{pmatrix} \begin{pmatrix} 0.390 & -0.731 \\ -0.731 & 1.48 \end{pmatrix} \begin{pmatrix} -50/\widehat{\beta}_3 \\ 50\widehat{\beta}_2/\widehat{\beta}_3^2 \end{pmatrix} \times 10^{-4}} \\
&= 7.0.
\end{aligned}
$$

The calculations show that the estimate of the percentage return to education (for married black women) is about 12% per year, with a standard error of 0.8. The estimate of the percentage return to experience for those with 10 years of experience is 1.2% per year, with a standard error of 0.6. And the estimate of the experience level which maximizes expected log wages is 35 years, with a standard error of 7.

7.12 t-statistic

Let $\theta = r(\boldsymbol{\beta}) : \mathbb{R}^k \to \mathbb{R}$ be a parameter of interest, $\widehat{\theta}$ its estimate and $s(\widehat{\theta})$ its asymptotic standard error. Consider the statistic

$$
T(\theta) = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})}. \tag{7.43}
$$

Different writers have called (7.43) a **t-statistic**, a **t-ratio**, a **z-statistic** or a **studentized statistic**, sometimes using the different labels to distinguish between finite-sample and asymptotic inference. As the statistics themselves are always (7.43) we won't make this distinction, and will simply refer to $T(\theta)$ as a t-statistic or a t-ratio. We also often suppress the parameter dependence, writing it as T. The t-statistic is a simple function of the estimate, its standard error, and the parameter.

By Theorems 7.10.2 and 7.10.3, $\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathrm{N}(0, V_\theta)$ and $\widehat{V}_\theta \xrightarrow{p} V_\theta$. Thus

$$
\begin{aligned}
T(\theta) &= \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \\
&= \frac{\sqrt{n}\left(\widehat{\theta} - \theta\right)}{\sqrt{\widehat{V}_\theta}} \\
&\xrightarrow{d} \frac{\mathrm{N}(0, V_\theta)}{\sqrt{V_\theta}} \\
&= \mathrm{Z} \sim \mathrm{N}(0, 1).
\end{aligned}
$$

The last equality is by the property that affine functions of normal distributions are normal (Theorem 5.2.3).

Thus the asymptotic distribution of the t-ratio $T(\theta)$ is the standard normal. Since this distribution does not depend on the parameters, we say that $T(\theta)$ is **asymptotically pivotal**. In finite samples $T(\theta)$ is not necessarily pivotal (as in the normal regression model) but the property states that the dependence on unknowns diminishes as n increases.

As we will see in the next section, it is also useful to consider the distribution of the **absolute t-ratio** $|T(\theta)|$. Since $T(\theta) \xrightarrow{d} \mathrm{Z}$, the continuous mapping theorem yields $|T(\theta)| \xrightarrow{d} |\mathrm{Z}|$. Letting $\Phi(u) = \Pr(\mathrm{Z} \leq u)$ denote the standard normal distribution function, we can calculate that the distribution function of $|\mathrm{Z}|$ is

$$\begin{aligned} \Pr(|\mathrm{Z}| \leq u) &= \Pr(-u \leq \mathrm{Z} \leq u) \\ &= \Pr(\mathrm{Z} \leq u) - \Pr(\mathrm{Z} < -u) \\ &= \Phi(u) - \Phi(-u) \\ &= 2\Phi(u) - 1 \end{aligned} \tag{7.44}$$

Theorem 7.12.1 *Under Assumptions 7.1.2 and 7.10.1,* $T(\theta) \xrightarrow{d} \mathrm{Z} \sim \mathrm{N}(0,1)$ *and* $|t_n(\theta)| \xrightarrow{d} |\mathrm{Z}|$.

The asymptotic normality of Theorem 7.12.1 is used to justify confidence intervals and tests for the parameters.

7.13 Confidence Intervals

The estimate $\widehat{\boldsymbol{\theta}}$ is a **point estimate** for $\boldsymbol{\theta}$, meaning that $\widehat{\boldsymbol{\theta}}$ is a single value in $\mathbb{R}^q$. A broader concept is a **set estimate** $\widehat{C}$ which is a collection of values in $\mathbb{R}^q$. When the parameter θ is real-valued then it is common to focus on sets of the form $\widehat{C} = [\widehat{L}, \widehat{U}]$ which is called an **interval estimate** for θ.

An interval estimate $\widehat{C}$ is a function of the data and hence is random. The **coverage probability** of the interval $\widehat{C} = [\widehat{L}, \widehat{U}]$ is $\Pr(\theta \in \widehat{C})$. The randomness comes from $\widehat{C}$ as the parameter θ is treated as fixed. In Section 5.12 we introduced confidence intervals for the normal regression model, which used the finite sample distribution of the t-statistic to construct exact confidence intervals for the regression coefficients. When we are outside the normal regression model we cannot rely on the exact normal distribution theory, but instead use asymptotic approximations. A benefit is that we can construct confidence intervals for general parameters of interest θ, not just regression coefficients.

An interval estimate $\widehat{C}$ is called a **confidence interval** when the goal is to set the coverage probability to equal a pre-specified target such as 90% or 95%. $\widehat{C}$ is called a $1-\alpha$ confidence interval if $\inf_\theta \Pr_\theta(\theta \in \widehat{C}) = 1-\alpha$.

When $\widehat{\theta}$ is asymptotically normal with standard error $s(\widehat{\theta})$, the conventional confidence interval for θ takes the form

$$\widehat{C} = \left[\widehat{\theta} - c \cdot s(\widehat{\theta}), \quad \widehat{\theta} + c \cdot s(\widehat{\theta})\right] \tag{7.45}$$

where c equals the $1-\alpha$ quantile of the distribution of $|\mathrm{Z}|$. Using (7.44) we calculate that c is equivalently the $1-\alpha/2$ quantile of the standard normal distribution. Thus, c solves

$$2\Phi(c) - 1 = 1 - \alpha.$$

This can be computed by, for example, `norminv(1-α/2)` in MATLAB. The confidence interval (7.45) is symmetric about the point estimate $\widehat{\theta}$, and its length is proportional to the standard error $s(\widehat{\theta})$.

Equivalently, (7.45) is the set of parameter values for θ such that the t-statistic $T(\theta)$ is smaller (in absolute value) than c, that is

$$\widehat{C} = \{\theta : |T(\theta)| \leq c\} = \left\{\theta : -c \leq \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \leq c\right\}.$$

The coverage probability of this confidence interval is

$$\Pr\left(\theta \in \widehat{C}\right) = \Pr\left(|T(\theta)| \leq c\right) \to \Pr\left(|\mathrm{Z}| \leq c\right) = 1 - \alpha$$

where the limit is taken as $n \to \infty$, and holds since $T(\theta)$ is asymptotically |Z| by Theorem 7.12.1. We call the limit the **asymptotic coverage probability**, and call $\widehat{C}$ an asymptotic $1-\alpha\%$ confidence interval for θ. Since the t-ratio is asymptotically pivotal, the asymptotic coverage probability is independent of the parameter θ.

It is useful to contrast the confidence interval (7.45) with (5.12) for the normal regression model. They are similar, but there are differences. The normal regression interval (5.12) only applies to regression coefficients β, not to functions θ of the coefficients. The normal interval (5.12) also is constructed with the homoskedastic standard error, while (7.45) can be constructed with a heteroskedastic-robust standard error. Furthermore, the constants c in (5.12) are calculated using the student t distribution, while c in (7.45) are calculated using the normal distribution. The difference between the student t and normal values are typically small in practice (since sample sizes are large in typical economic applications). However, since the student t values are larger, it results in slightly larger confidence intervals, which is probably reasonable. (A practical rule of thumb is that if the sample sizes are sufficiently small that it makes a difference, then probably neither (5.12) nor (7.45) should be trusted.) Despite these differences, the coincidence of the intervals means that inference on regression coefficients is generally robust to using either the exact normal sampling assumption or the asymptotic large sample approximation, at least in large samples.

In Stata, by default the program reports 95% confidence intervals for each coefficient where the critical values c are calculated using the t_{n-k} distribution. This is done for all standard error methods even though it is only justified for homoskedastic standard errors and under normality.

The standard coverage probability for confidence intervals is 95%, leading to the choice $c = 1.96$ for the constant in (7.45). Rounding 1.96 to 2, we obtain what might be the most commonly used confidence interval in applied econometric practice

$$\widehat{C} = \left[\widehat{\theta} - 2s(\widehat{\theta}), \quad \widehat{\theta} + 2s(\widehat{\theta})\right]. \tag{7.46}$$

This is a useful rule-of thumb. This asymptotic 95% confidence interval $\widehat{C}$ is simple to compute and can be roughly calculated from tables of coefficient estimates and standard errors. (Technically, it is an asymptotic 95.4% interval, due to the substitution of 2.0 for 1.96, but this distinction is overly precise.)

Theorem 7.13.1 *Under Assumptions 7.1.2 and 7.10.1, for $\widehat{C}$ defined in (7.45), with $c = \Phi^{-1}(1 - \alpha/2)$, $\Pr\left(\theta \in \widehat{C}\right) \longrightarrow 1 - \alpha$. For $c = 1.96$, $\Pr\left(\theta \in \widehat{C}\right) \longrightarrow 0.95$.*

Confidence intervals are a simple yet effective tool to assess estimation uncertainty. When reading a set of empirical results, look at the estimated coefficient estimates and the standard errors. For a parameter of interest, compute the confidence interval C_n and consider the meaning of the spread of the suggested values. If the range of values in the confidence interval are too wide to learn about θ, then do not jump to a conclusion about θ based on the point estimate alone.

For illustration, consider the three examples presented in Section 7.11 based on the log wage regression for married black women.

Percentage return to education. A 95% asymptotic confidence interval is $11.8 \pm 1.96 \times 0.8 = [10.2, 13.3]$.

Percentage return to experience for individuals with 10 years experience. A 90% asymptotic confidence interval is $1.1 \pm 1.645 \times 0.4 = [0.5, 1.8]$.

Experience level which maximizes expected log wages. An 80% asymptotic confidence interval is $35 \pm 1.28 \times 7 = [26, 44]$.

7.14 Regression Intervals

In the linear regression model the conditional mean of y_i given $\boldsymbol{x}_i = \boldsymbol{x}$ is

$$m(\boldsymbol{x}) = \mathbb{E}\left(y_i \mid \boldsymbol{x}_i = \boldsymbol{x}\right) = \boldsymbol{x}'\boldsymbol{\beta}.$$

In some cases, we want to estimate $m(\boldsymbol{x})$ at a particular point $\boldsymbol{x}$. Notice that this is a linear function of $\boldsymbol{\beta}$. Letting $r(\boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$ and $\theta = r(\boldsymbol{\beta})$, we see that $\widehat{m}(\boldsymbol{x}) = \widehat{\theta} = \boldsymbol{x}'\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{R} = \boldsymbol{x}$, so $s(\widehat{\theta}) = \sqrt{\boldsymbol{x}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}}$. Thus an asymptotic 95% confidence interval for $m(\boldsymbol{x})$ is

$$\left[\boldsymbol{x}'\widehat{\boldsymbol{\beta}} \pm 1.96\sqrt{\boldsymbol{x}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}}\right].$$

It is interesting to observe that if this is viewed as a function of $\boldsymbol{x}$, the width of the confidence set is dependent on $\boldsymbol{x}$.

To illustrate, we return to the log wage regression (3.14) of Section 3.7. The estimated regression equation is

$$\widehat{\log(Wage)} = \boldsymbol{x}'\widehat{\boldsymbol{\beta}} = 0.155x + 0.698$$

where $x = education$. The covariance matrix estimate from (4.44) is

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \begin{pmatrix} 0.001 & -0.015 \\ -0.015 & 0.243 \end{pmatrix}.$$

Thus the 95% confidence interval for the regression takes the form

$$0.155x + 0.698 \pm 1.96\sqrt{0.001x^2 - 0.030x + 0.243}.$$

The estimated regression and 95% intervals are shown in Figure 7.6. Notice that the confidence bands take a hyperbolic shape. This means that the regression line is less precisely estimated for very large and very small values of *education*.

Plots of the estimated regression line and confidence intervals are especially useful when the regression includes nonlinear terms. To illustrate, consider the log wage regression (7.41) which includes experience and its square, with covariance matrix (7.42). We are interested in plotting the regression estimate and regression intervals as a function of *experience*. Since the regression also includes *education,* to plot the estimates in a simple graph we need to fix *education* at a

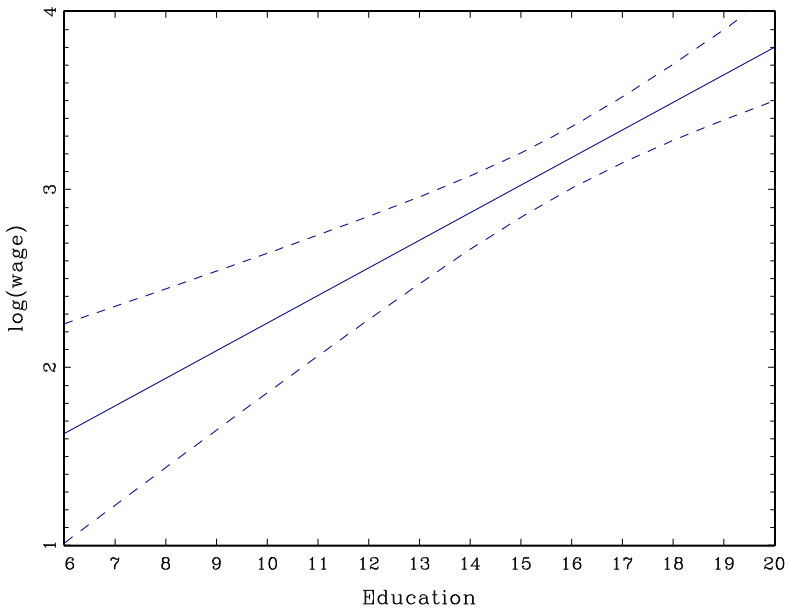


Figure 7.6: Wage on Education Regression Intervals

specific value. We select *education*=12. This only affects the level of the estimated regression, since *education* enters without an interaction. Define the points of evaluation

$$\boldsymbol{z}(x) = \begin{pmatrix} 12 \\ x \\ x^2/100 \\ 1 \end{pmatrix}$$

where $x =$*experience*.

Thus the 95% regression interval for *education*=12, as a function of $x =$*experience* is

$$\begin{aligned} &0.118 \times 12 \ + 0.016\ x - 0.022\ x^2/100 + 0.947 \\ &\pm 1.96 \sqrt{\boldsymbol{z}(x)' \begin{pmatrix} 0.632 & 0.131 & -0.143 & -11.1 \\ 0.131 & 0.390 & -0.731 & -6.25 \\ -0.143 & -0.731 & 1.48 & 9.43 \\ -11.1 & -6.25 & 9.43 & 246 \end{pmatrix} \boldsymbol{z}(x) \times 10^{-4}} \\ &= 0.016\ x - .00022\ x^2 + 2.36 \\ &\pm 0.0196 \sqrt{70.608 - 9.356\ x + 0.54428\ x^2 - 0.01462\ x^3 + 0.000148\ x^4}. \end{aligned}$$

The estimated regression and 95% intervals are shown in Figure 7.7. The regression interval widens greatly for small and large values of experience, indicating considerable uncertainty about the effect of experience on mean wages for this population. The confidence bands take a more complicated shape than in Figure 7.6 due to the nonlinear specification.

7.15 Forecast Intervals

Suppose we are given a value of the regressor vector $\boldsymbol{x}_{n+1}$ for an individual outside the sample, and we want to forecast (guess) y_{n+1} for this individual. This is equivalent to forecasting y_{n+1} given $\boldsymbol{x}_{n+1} = \boldsymbol{x}$, which will generally be a function of $\boldsymbol{x}$. A reasonable forecasting rule is the conditional mean $m(\boldsymbol{x})$ as it is the mean-square-minimizing forecast. A point forecast is the estimated conditional mean $\widehat{m}(\boldsymbol{x}) = \boldsymbol{x}'\widehat{\boldsymbol{\beta}}$. We would also like a measure of uncertainty for the forecast.

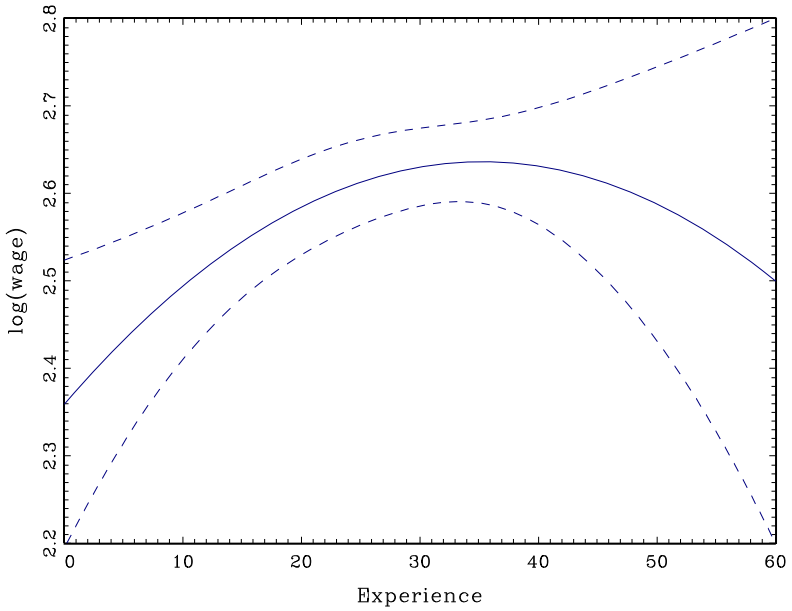


Figure 7.7: Wage on Experience Regression Intervals

The forecast error is $\widehat{e}_{n+1} = y_{n+1} - \widehat{m}(\boldsymbol{x}) = e_{n+1} - \boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$. As the out-of-sample error e_{n+1} is independent of the in-sample estimate $\widehat{\boldsymbol{\beta}}$, this has conditional variance

$$\begin{aligned}
\mathbb{E}\left(\widehat{e}_{n+1}^2 | \boldsymbol{x}_{n+1} = \boldsymbol{x}\right) &= \mathbb{E}\left(e_{n+1}^2 - 2\boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) e_{n+1} + \boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\boldsymbol{x} | \boldsymbol{x}_{n+1} = \boldsymbol{x}\right) \\
&= \mathbb{E}\left(e_{n+1}^2 \mid \boldsymbol{x}_{n+1} = \boldsymbol{x}\right) + \boldsymbol{x}'\mathbb{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{x} \\
&= \sigma^2(\boldsymbol{x}) + \boldsymbol{x}'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}.
\end{aligned}$$

Under homoskedasticity $\mathbb{E}\left(e_{n+1}^2 \mid \boldsymbol{x}_{n+1}\right) = \sigma^2$, the natural estimate of this variance is $\widehat{\sigma}^2 + \boldsymbol{x}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}$, so a standard error for the forecast is $\widehat{s}(\boldsymbol{x}) = \sqrt{\widehat{\sigma}^2 + \boldsymbol{x}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}}$. Notice that this is different from the standard error for the conditional mean.

The conventional 95% forecast interval for y_{n+1} uses a normal approximation and sets

$$\left[\boldsymbol{x}'\widehat{\boldsymbol{\beta}} \pm 2\widehat{s}(\boldsymbol{x})\right].$$

It is difficult, however, to fully justify this choice. It would be correct if we have a normal approximation to the ratio

$$\frac{e_{n+1} - \boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{\widehat{s}(\boldsymbol{x})}.$$

The difficulty is that the equation error e_{n+1} is generally non-normal, and asymptotic theory cannot be applied to a single observation. The only special exception is the case where e_{n+1} has the exact distribution $\mathrm{N}(0, \sigma^2)$, which is generally invalid.

To get an accurate forecast interval, we need to estimate the conditional distribution of e_{n+1} given $\boldsymbol{x}_{n+1} = \boldsymbol{x}$, which is a much more difficult task. Perhaps due to this difficulty, many applied forecasters use the simple approximate interval $\left[\boldsymbol{x}'\widehat{\boldsymbol{\beta}} \pm 2\widehat{s}(\boldsymbol{x})\right]$ despite the lack of a convincing justification.

7.16 Wald Statistic

Let $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta}) : \mathbb{R}^k \to \mathbb{R}^q$ be any parameter vector of interest, $\widehat{\boldsymbol{\theta}}$ its estimate and $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}$ its covariance matrix estimator. Consider the quadratic form

$$W(\boldsymbol{\theta}) = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = n \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right). \tag{7.47}$$

where $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} = n\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}$. When $q = 1$, then $W(\theta) = T(\theta)^2$ is the square of the t-ratio. When $q > 1$, $W(\theta)$ is typically called a **Wald statistic**. We are interested in its sampling distribution.

The asymptotic distribution of $W(\theta)$ is simple to derive given Theorem 7.10.2 and Theorem 7.10.3, which show that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{Z} \sim \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

and

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\theta}}.$$

Note that $\boldsymbol{V}_{\boldsymbol{\theta}} > 0$ since $\boldsymbol{R}$ is full rank under Assumption 7.10.1. It follows that

$$W(\boldsymbol{\theta}) = \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{Z}' \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \mathrm{Z} \tag{7.48}$$

a quadratic in the normal random vector Z. As shown in Theorem 5.3.3, the distribution of this quadratic form is χ^2_q, a chi-square random variable with q degrees of freedom.

Theorem 7.16.1 *Under Assumptions 7.1.2 and 7.10.1, as $n \to \infty$,*

$$W(\boldsymbol{\theta}) \xrightarrow{d} \chi^2_q.$$

Theorem 7.16.1 is used to justify multivariate confidence regions and multivariate hypothesis tests.

7.17 Homoskedastic Wald Statistic

Under the conditional homoskedasticity assumption $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$ we can construct the Wald statistic using the homoskedastic covariance matrix estimator $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^0$ defined in (7.39). This yields a homoskedastic Wald statistic

$$W^0(\boldsymbol{\theta}) = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^0\right)^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = n \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \left(\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^0\right)^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right). \tag{7.49}$$

Under the additional assumption of conditional homoskedasticity, it has the same asymptotic distribution as $W(\theta)$.

Theorem 7.17.1 *Under Assumptions 7.1.2 and 7.10.1, and $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, as $n \to \infty$,*

$$W^0(\boldsymbol{\theta}) \xrightarrow{d} \chi^2_q.$$

7.18 Confidence Regions

A confidence region $\widehat{C}$ is a set estimator for $\boldsymbol{\theta} \in \mathbb{R}^q$ when $q > 1$. A confidence region $\widehat{C}$ is a set in $\mathbb{R}^q$ intended to cover the true parameter value with a pre-selected probability $1-\alpha$. Thus an ideal confidence region has the coverage probability $\Pr(\boldsymbol{\theta} \in \widehat{C}) = 1-\alpha$. In practice it is typically not possible to construct a region with exact coverage, but we can calculate its asymptotic coverage.

When the parameter estimate satisfies the conditions of Theorem 7.16.1, a good choice for a confidence region is the ellipse

$$\widehat{C} = \{\boldsymbol{\theta} : W(\boldsymbol{\theta}) \leq c_{1-\alpha}\}.$$

with $c_{1-\alpha}$ the $1-\alpha$ quantile of the χ^2_q distribution. (Thus $F_q(c_{1-\alpha}) = 1-\alpha$.) It can be computed by, for example, `chi2inv(1-`α`,q)`in MATLAB.

Theorem 7.16.1 implies

$$\Pr\left(\boldsymbol{\theta} \in \widehat{C}\right) \to \Pr\left(\chi^2_q \leq c_{1-\alpha}\right) = 1-\alpha$$

which shows that $\widehat{C}$ has asymptotic coverage $1-\alpha$.

To illustrate the construction of a confidence region, consider the estimated regression (7.41) of the model

$$\widehat{\log(Wage)} = \beta_1\, education + \beta_2\, experience + \beta_3\, experience^2/100 + \beta_4.$$

Suppose that the two parameters of interest are the percentage return to education $\theta_1 = 100\beta_1$ and the percentage return to experience for individuals with 10 years experience $\theta_2 = 100\beta_2 + 20\beta_3$. These two parameters are a linear transformation of the regression parameters with point estimates

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} 100 & 0 & 0 & 0 \\ 0 & 100 & 20 & 0 \end{pmatrix} \widehat{\boldsymbol{\beta}} = \begin{pmatrix} 11.8 \\ 1.2 \end{pmatrix},$$

and have the covariance matrix estimate

$$\begin{aligned} \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}} &= \begin{pmatrix} 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 20 \end{pmatrix} \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \begin{pmatrix} 0 & 0 \\ 100 & 0 \\ 0 & 100 \\ 0 & 20 \end{pmatrix} \\ &= \begin{pmatrix} 0.632 & 0.103 \\ 0.103 & 0.157 \end{pmatrix} \end{aligned}$$

with inverse

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} = \begin{pmatrix} 1.77 & -1.16 \\ -1.16 & 7.13 \end{pmatrix}.$$

Thus the Wald statistic is

$$\begin{aligned} W(\boldsymbol{\theta}) &= \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \\ &= \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix}' \begin{pmatrix} 1.77 & -1.16 \\ -1.16 & 7.13 \end{pmatrix} \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix} \\ &= 1.77\,(11.8 - \theta_1)^2 - 2.32\,(11.8 - \theta_1)\,(1.2 - \theta_2) + 7.13\,(1.2 - \theta_2)^2. \end{aligned}$$

The 90% quantile of the χ^2_2 distribution is 4.605 (we use the χ^2_2 distribution as the dimension of $\boldsymbol{\theta}$ is two), so an asymptotic 90% confidence region for the two parameters is the interior of the ellipse $W(\boldsymbol{\theta}) = 4.605$ which is displayed in Figure 7.8. Since the estimated correlation of the two coefficient estimates is modest (about 0.3) the region is modestly elliptical.

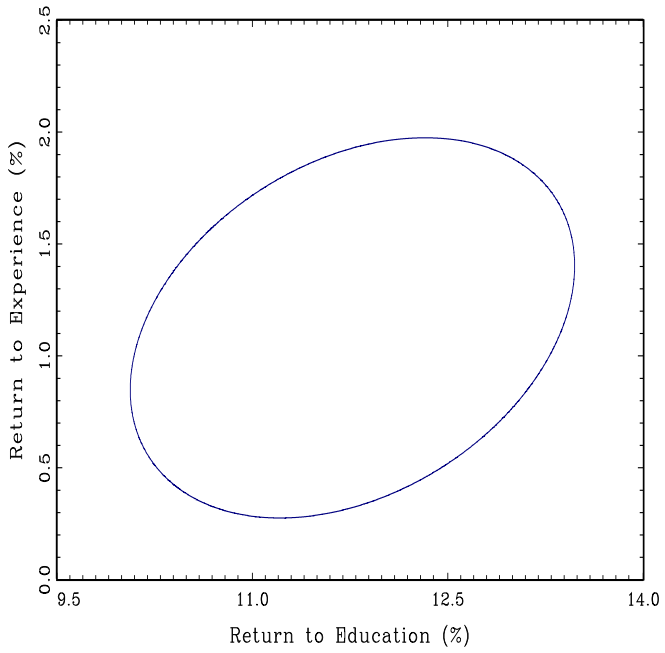


Figure 7.8: Confidence Region for Return to Experience and Return to Education

7.19 Semiparametric Efficiency in the Projection Model

In Section 4.7 we presented the Gauss-Markov theorem, which stated that in the homoskedastic CEF model, in the class of linear unbiased estimators the one with the smallest variance is least-squares. As we noted in that section, the restriction to linear unbiased estimators is unsatisfactory as it leaves open the possibility that an alternative (non-linear) estimator could have a smaller asymptotic variance. In addition, the restriction to the homoskedastic CEF model is also unsatisfactory as the projection model is more relevant for empirical application. The question remains: what is the most efficient estimator of the projection coefficient $\boldsymbol{\beta}$ (or functions $\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\beta})$) in the projection model?

It turns out that it is straightforward to show that the projection model falls in the estimator class considered in Proposition 6.15.2. It follows that the least-squares estimator is semiparametrically efficient in the sense that it has the smallest asymptotic variance in the class of semiparametric estimators of $\boldsymbol{\beta}$. This is a more powerful and interesting result than the Gauss-Markov theorem.

To see this, it is worth rephrasing Proposition 6.15.2 with amended notation. Suppose that a parameter of interest is $\boldsymbol{\theta} = \boldsymbol{g}(\boldsymbol{\mu})$ where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{z}_i)$, for which the moment estimators are $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i$ and $\widehat{\boldsymbol{\theta}} = \boldsymbol{g}(\widehat{\boldsymbol{\mu}})$. Let

$$\mathcal{L}_2(\boldsymbol{g}) = \left\{ F : \mathbb{E}\left\|\boldsymbol{z}\right\|^2 < \infty,\ \boldsymbol{g}(\boldsymbol{u}) \text{ is continuously differentiable at } \boldsymbol{u} = \mathbb{E}(\boldsymbol{z}) \right\}$$

be the set of distributions for which $\widehat{\boldsymbol{\theta}}$ satisfies the central limit theorem.

Proposition 7.19.1 *In the class of distributions $F \in \mathcal{L}_2(\boldsymbol{g})$, $\widehat{\boldsymbol{\theta}}$ is semiparametrically efficient for $\boldsymbol{\theta}$ in the sense that its asymptotic variance equals the semiparametric efficiency bound.*

Proposition 7.19.1 says that under the minimal conditions in which $\widehat{\boldsymbol{\theta}}$ is asymptotically normal, then no semiparametric estimator can have a smaller asymptotic variance than $\widehat{\boldsymbol{\theta}}$.

To show that an estimator is semiparametrically efficient it is sufficient to show that it falls in the class covered by this Proposition. To show that the projection model falls in this class, we write $\boldsymbol{\beta} = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xy}} = \boldsymbol{g}(\boldsymbol{\mu})$ where $\boldsymbol{\mu} = \mathbb{E}(\boldsymbol{z}_i)$ and $\boldsymbol{z}_i = (\boldsymbol{x}_i\boldsymbol{x}_i', \boldsymbol{x}_i y_i)$. The class $\mathcal{L}_2(\boldsymbol{g})$ equals the class of distributions

$$\mathcal{L}_4(\boldsymbol{\beta}) = \left\{F : \mathbb{E}\left(y^4\right) < \infty,\ \mathbb{E}\left\|\boldsymbol{x}\right\|^4 < \infty,\ \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) > 0\right\}.$$

Proposition 7.19.2 *In the class of distributions $F \in \mathcal{L}_4(\boldsymbol{\beta})$, the least-squares estimator $\widehat{\boldsymbol{\beta}}$ is semiparametrically efficient for $\boldsymbol{\beta}$.*

The least-squares estimator is an asymptotically efficient estimator of the projection coefficient because the latter is a smooth function of sample moments and the model implies no further restrictions. However, if the class of permissible distributions is restricted to a strict subset of $\mathcal{L}_4(\boldsymbol{\beta})$ then least-squares can be inefficient. For example, the linear CEF model with heteroskedastic errors is a strict subset of $\mathcal{L}_4(\boldsymbol{\beta})$, and the GLS estimator has a smaller asymptotic variance than OLS. In this case, the knowledge that true conditional mean is linear allows for more efficient estimation of the unknown parameter.

From Proposition 7.19.1 we can also deduce that plug-in estimators $\widehat{\boldsymbol{\theta}} = \boldsymbol{h}(\widehat{\boldsymbol{\beta}})$ are semiparametrically efficient estimators of $\boldsymbol{\theta} = \boldsymbol{h}(\boldsymbol{\beta})$ when $\boldsymbol{h}$ is continuously differentiable. We can also deduce that other parameters estimators are semiparametrically efficient, such as $\widehat{\sigma}^2$ for σ^2. To see this, note that we can write

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left(\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right) \\
&= \mathbb{E}\left(y_i^2\right) - 2\mathbb{E}\left(y_i\boldsymbol{x}_i'\right)\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\boldsymbol{\beta} \\
&= Q_{yy} - \boldsymbol{Q}_{\boldsymbol{yx}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xy}}
\end{aligned}$$

which is a smooth function of the moments Q_{yy}, $\boldsymbol{Q}_{\boldsymbol{yx}}$ and $\boldsymbol{Q}_{\boldsymbol{xx}}$. Similarly the estimator $\widehat{\sigma}^2$ equals

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 \\
&= \widehat{Q}_{yy} - \widehat{\boldsymbol{Q}}_{\boldsymbol{yx}}\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\widehat{\boldsymbol{Q}}_{\boldsymbol{xy}}.
\end{aligned}$$

Since the variables y_i^2, $y_i\boldsymbol{x}_i'$ and $\boldsymbol{x}_i\boldsymbol{x}_i'$ all have finite variances when $F \in \mathcal{L}_4(\boldsymbol{\beta})$, the conditions of Proposition 7.19.1 are satisfied. We conclude:

Proposition 7.19.3 *In the class of distributions $F \in \mathcal{L}_4(\boldsymbol{\beta})$, $\widehat{\sigma}^2$ is semiparametrically efficient for σ^2.*

7.20 Semiparametric Efficiency in the Homoskedastic Regression Model*

In Section 7.19 we showed that the OLS estimator is semiparametrically efficient in the projection model. What if we restrict attention to the classical homoskedastic regression model? Is OLS still efficient in this class? In this section we derive the asymptotic semiparametric efficiency bound

for this model, and show that it is the same as that obtained by the OLS estimator. Therefore it turns out that least-squares is efficient in this class as well.

Recall that in the homoskedastic regression model the asymptotic variance of the OLS estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ is $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\sigma^2$. Therefore, as described in Section 6.15, it is sufficient to find a parametric submodel whose Cramer-Rao bound for estimation of $\boldsymbol{\beta}$ is $\boldsymbol{V}_{\boldsymbol{\beta}}^0$. This would establish that $\boldsymbol{V}_{\boldsymbol{\beta}}^0$ is the semiparametric variance bound and the OLS estimator $\widehat{\boldsymbol{\beta}}$ is semiparametrically efficient for $\boldsymbol{\beta}$.

Let the joint density of y and $\boldsymbol{x}$ be written as $f(y, \boldsymbol{x}) = f_1(y \mid \boldsymbol{x}) f_2(\boldsymbol{x})$, the product of the conditional density of y given $\boldsymbol{x}$ and the marginal density of $\boldsymbol{x}$. Now consider the parametric submodel

$$f(y, \boldsymbol{x} \mid \boldsymbol{\theta}) = f_1(y \mid \boldsymbol{x}) \left(1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) f_2(\boldsymbol{x}). \tag{7.50}$$

You can check that in this submodel the marginal density of $\boldsymbol{x}$ is $f_2(\boldsymbol{x})$ and the conditional density of y given $\boldsymbol{x}$ is $f_1(y \mid \boldsymbol{x})\left(1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right)$. To see that the latter is a valid conditional density, observe that the regression assumption implies that $\int y f_1(y \mid \boldsymbol{x})\, dy = \boldsymbol{x}'\boldsymbol{\beta}$ and therefore

$$\begin{aligned}
&\int f_1(y \mid \boldsymbol{x}) \left(1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) dy \\
&= \int f_1(y \mid \boldsymbol{x})\, dy + \int f_1(y \mid \boldsymbol{x}) \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right) dy \left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2 \\
&= 1.
\end{aligned}$$

In this parametric submodel the conditional mean of y given $\boldsymbol{x}$ is

$$\begin{aligned}
\mathbb{E}_\theta(y \mid \boldsymbol{x}) &= \int y f_1(y \mid \boldsymbol{x}) \left(1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) dy \\
&= \int y f_1(y \mid \boldsymbol{x})\, dy + \int y f_1(y \mid \boldsymbol{x}) \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2 dy \\
&= \int y f_1(y \mid \boldsymbol{x})\, dy + \int \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2 f_1(y \mid \boldsymbol{x}) \left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2 dy \\
&+ \int \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right) f_1(y \mid \boldsymbol{x})\, dy \left(\boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2 \\
&= \boldsymbol{x}'(\boldsymbol{\beta} + \boldsymbol{\theta}),
\end{aligned}$$

using the homoskedasticity assumption $\int \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^2 f_1(y \mid \boldsymbol{x})\, dy = \sigma^2$. This means that in this parametric submodel, the conditional mean is linear in $\boldsymbol{x}$ and the regression coefficient is $\boldsymbol{\beta}(\boldsymbol{\theta}) = \boldsymbol{\beta} + \boldsymbol{\theta}$.

We now calculate the score for estimation of $\boldsymbol{\theta}$. Since

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(y, \boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) = \frac{\boldsymbol{x}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)/\sigma^2}{1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2}$$

the score is

$$\boldsymbol{s} = \frac{\partial}{\partial \boldsymbol{\theta}} \log f(y, \boldsymbol{x} \mid \boldsymbol{\theta}_0) = \boldsymbol{x}e/\sigma^2.$$

The Cramer-Rao bound for estimation of $\boldsymbol{\theta}$ (and therefore $\boldsymbol{\beta}(\boldsymbol{\theta})$ as well) is

$$\left(\mathbb{E}\left(\boldsymbol{s}\boldsymbol{s}'\right)\right)^{-1} = \left(\sigma^{-4}\mathbb{E}\left((\boldsymbol{x}e)(\boldsymbol{x}e)'\right)\right)^{-1} = \sigma^2 \boldsymbol{Q}_{\boldsymbol{xx}}^{-1} = \boldsymbol{V}_{\boldsymbol{\beta}}^0.$$

We have shown that there is a parametric submodel (7.50) whose Cramer-Rao bound for estimation of $\boldsymbol{\beta}$ is identical to the asymptotic variance of the least-squares estimator, which therefore is the semiparametric variance bound.

Theorem 7.20.1 *In the homoskedastic regression model, the semiparametric variance bound for estimation of $\boldsymbol{\beta}$ is $\mathbf{V}^0 = \sigma^2 \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$ and the OLS estimator is semiparametrically efficient.*

This result is similar to the Gauss-Markov theorem, in that it asserts the efficiency of the least-squares estimator in the context of the homoskedastic regression model. The difference is that the Gauss-Markov theorem states that OLS has the smallest variance among the set of unbiased linear estimators, while Theorem 7.20.1 states that OLS has the smallest asymptotic variance among all regular estimators. This is a much more powerful statement.

7.21 Uniformly Consistent Residuals*

It seems natural to view the residuals $\widehat{e}_i$ as estimates of the unknown errors e_i. Are they consistent estimates? In this section we develop an appropriate convergence result. This is not a widely-used technique, and can safely be skipped by most readers.

Notice that we can write the residual as

$$\begin{aligned} \widehat{e}_i &= y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} \\ &= e_i + \boldsymbol{x}_i' \boldsymbol{\beta} - x_i' \widehat{\boldsymbol{\beta}} \\ &= e_i - \boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right). \end{aligned} \tag{7.51}$$

Since $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$ it seems reasonable to guess that $\widehat{e}_i$ will be close to e_i if n is large.

We can bound the difference in (7.51) using the Schwarz inequality (A.20) to find

$$|\widehat{e}_i - e_i| = \left| \boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \right| \leq \|\boldsymbol{x}_i\| \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\|. \tag{7.52}$$

To bound (7.52) we can use $\left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\| = O_p(n^{-1/2})$ from Theorem 7.3.2, but we also need to bound the random variable $\|\boldsymbol{x}_i\|$. If the regressor is bounded, that is, $\|\boldsymbol{x}_i\| \leq B < \infty$, then $|\widehat{e}_i - e_i| \leq B \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\| = O_p(n^{-1/2})$. However if the regressor does not have bounded support then we have to be more careful.

The key is Theorem 6.14.1 which shows that $\mathbb{E} \|\boldsymbol{x}_i\|^r < \infty$ implies $\boldsymbol{x}_i = o_p\left(n^{1/r}\right)$ uniformly in i, or

$$n^{-1/r} \max_{1 \leq i \leq n} \|\boldsymbol{x}_i\| \xrightarrow{p} 0.$$

Applied to (7.52) we obtain

$$\begin{aligned} \max_{1 \leq i \leq n} |\widehat{e}_i - e_i| &\leq \max_{1 \leq i \leq n} \|\boldsymbol{x}_i\| \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\| \\ &= o_p(n^{-1/2+1/r}). \end{aligned}$$

We have shown the following.

Theorem 7.21.1 *Under Assumption 7.1.2 and $\mathbb{E} \|\boldsymbol{x}_i\|^r < \infty$, then uniformly in $1 \leq i \leq n$*

$$\widehat{e}_i = e_i + o_p(n^{-1/2+1/r}). \tag{7.53}$$

The rate of convergence in (7.53) depends on r. Assumption 7.1.2 requires $r \geq 4$, so the rate of convergence is at least $o_p(n^{-1/4})$. As r increases, the rate improves. As a limiting case, from Theorem 6.14.1 we see that if $\mathbb{E}\left(\exp(\boldsymbol{t}'\boldsymbol{x}_i)\right) < \infty$ for some $\boldsymbol{t} \neq 0$ then $\boldsymbol{x}_i = o_p\left((\log n)^{1+\eta}\right)$ uniformly in i, and thus $\widehat{e}_i = e_i + o_p\left(n^{-1/2}(\log n)^{1+\eta}\right)$.

We mentioned in Section 7.7 that there are multiple ways to prove the consistent of the covariance matrix estimator $\widehat{\boldsymbol{\Omega}}$. We now show that Theorem 7.21.1 provides one simple method to establish (7.31) and thus Theorem 7.7.1. Let $q_n = \max_{1\leq i\leq n} |\widehat{e}_i - e_i| = o_p(n^{-1/4})$. Since

$$\widehat{e}_i^2 - e_i^2 = 2e_i\left(\widehat{e}_i - e_i\right) + \left(\widehat{e}_i - e_i\right)^2,$$

then

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\left(\widehat{e}_i^2 - e_i^2\right)\right\| &\leq \frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{x}_i\boldsymbol{x}_i'\right\| \left|\widehat{e}_i^2 - e_i^2\right| \\
&\leq \frac{2}{n}\sum_{i=1}^n \left\|\boldsymbol{x}_i\right\|^2 |e_i|\,|\widehat{e}_i - e_i| + \frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{x}_i\right\|^2 |\widehat{e}_i - e_i|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^n \left\|\boldsymbol{x}_i\right\|^2 |e_i|\, q_n + \frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{x}_i\right\|^2 q_n^2 \\
&\leq o_p(n^{-1/4}).
\end{aligned}$$

7.22 Asymptotic Leverage*

Recall the definition of leverage from (3.25)

$$h_{ii} = \boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i.$$

These are the diagonal elements of the projection matrix $\boldsymbol{P}$ and appear in the formula for leave-one-out prediction errors and several covariance matrix estimators. We can show that under iid sampling the leverage values are uniformly asymptotically small.

Let $\lambda_{\min}(\boldsymbol{A})$ and $\lambda_{\max}(\boldsymbol{A})$ denote the smallest and largest eigenvalues of a symmetric square matrix $\boldsymbol{A}$, and note that $\lambda_{\max}(\boldsymbol{A}^{-1}) = (\lambda_{\min}(\boldsymbol{A}))^{-1}$.

Since $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{Q}_{\boldsymbol{xx}} > 0$ then by the CMT, $\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right) \xrightarrow{p} \lambda_{\min}\left(\boldsymbol{Q}_{\boldsymbol{xx}}\right) > 0$. (The latter is positive since $\boldsymbol{Q}_{\boldsymbol{xx}}$ is positive definite and thus all its eigenvalues are positive.) Then by the Quadratic Inequality (A.28)

$$\begin{aligned}
h_{ii} &= \boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i \\
&\leq \lambda_{\max}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(\boldsymbol{x}_i'\boldsymbol{x}_i\right) \\
&= \left(\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)\right)^{-1}\frac{1}{n}\left\|\boldsymbol{x}_i\right\|^2 \\
&\leq \left(\lambda_{\min}\left(\boldsymbol{Q}_{\boldsymbol{xx}}\right) + o_p(1)\right)^{-1}\frac{1}{n}\max_{1\leq i\leq n}\left\|\boldsymbol{x}_i\right\|^2. \qquad (7.54)
\end{aligned}$$

Theorem 6.14.1 shows that $\mathbb{E}\left\|\boldsymbol{x}_i\right\|^r < \infty$ implies $\max_{1\leq i\leq n}\left\|\boldsymbol{x}_i\right\|^2 = \left(\max_{1\leq i\leq n}\left\|\boldsymbol{x}_i\right\|\right)^2 = o_p\left(n^{2/r}\right)$ and thus (7.54) is $o_p\left(n^{2/r-1}\right)$.

Theorem 7.22.1 *If $\boldsymbol{x}_i$ is independent and identically distributed and $\mathbb{E}\left\|\boldsymbol{x}_i\right\|^r < \infty$ for some $r \geq 2$, then uniformly in $1 \leq i \leq n$, $h_{ii} = o_p\left(n^{2/r-1}\right)$.*

For any $r \geq 2$ then $h_{ii} = o_p(1)$ (uniformly in $i \leq n$). Larger r implies a stronger rate of convergence, for example $r = 4$ implies $h_{ii} = o_p\left(n^{-1/2}\right)$.

Theorem (7.22.1) implies that under random sampling with finite variances and large samples, no individual observation should have a large leverage value. Consequently individual observations should not be influential, unless one of these conditions is violated.

Exercises

Exercise 7.1 Take the model $y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i$ with $\mathbb{E}(\boldsymbol{x}_i e_i) = \mathbf{0}$. Suppose that $\boldsymbol{\beta}_1$ is estimated by regressing y_i on $\boldsymbol{x}_{1i}$ only. Find the probability limit of this estimator. In general, is it consistent for $\boldsymbol{\beta}_1$? If not, under what conditions is this estimator consistent for $\boldsymbol{\beta}_1$?

Exercise 7.2 Let $\boldsymbol{y}$ be $n \times 1$, $\boldsymbol{X}$ be $n \times k$ (rank k). $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with $\mathbb{E}(\boldsymbol{x}_i e_i) = 0$. Define the *ridge regression* estimator

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' + \lambda \boldsymbol{I}_k\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \tag{7.55}$$

here $\lambda > 0$ is a fixed constant. Find the probability limit of $\widehat{\boldsymbol{\beta}}$ as $n \to \infty$. Is $\widehat{\boldsymbol{\beta}}$ consistent for $\boldsymbol{\beta}$?

Exercise 7.3 For the ridge regression estimator (7.55), set $\lambda = cn$ where $c > 0$ is fixed as $n \to \infty$. Find the probability limit of $\widehat{\boldsymbol{\beta}}$ as $n \to \infty$.

Exercise 7.4 Verify some of the calculations reported in Section 7.4. Specifically, suppose that x_{1i} and x_{2i} only take the values $\{-1, +1\}$, symmetrically, with

$$\begin{aligned} \Pr(x_{1i} = x_{2i} = 1) &= \Pr(x_{1i} = x_{2i} = -1) = 3/8 \\ \Pr(x_{1i} = 1, x_{2i} = -1) &= \Pr(x_{1i} = -1, x_{2i} = 1) = 1/8 \\ \mathbb{E}\left(e_i^2 \mid x_{1i} = x_{2i}\right) &= \frac{5}{4} \\ \mathbb{E}\left(e_i^2 \mid x_{1i} \neq x_{2i}\right) &= \frac{1}{4}. \end{aligned}$$

Verify the following:

(a) $\mathbb{E}(x_{1i}) = 0$

(b) $\mathbb{E}\left(x_{1i}^2\right) = 1$

(c) $\mathbb{E}(x_{1i}x_{2i}) = \dfrac{1}{2}$

(d) $\mathbb{E}\left(e_i^2\right) = 1$

(e) $\mathbb{E}\left(x_{1i}^2 e_i^2\right) = 1$

(f) $\mathbb{E}\left(x_{1i}x_{2i}e_i^2\right) = \dfrac{7}{8}$.

Exercise 7.5 Show (7.19)-(7.22).

Exercise 7.6 The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= \mathbf{0} \\ \boldsymbol{\Omega} &= \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right). \end{aligned}$$

Find the method of moments estimators $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Omega}})$ for $(\boldsymbol{\beta}, \boldsymbol{\Omega})$.

(a) In this model, are $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\Omega}})$ efficient estimators of $(\boldsymbol{\beta}, \boldsymbol{\Omega})$?

(b) If so, in what sense are they efficient?

Exercise 7.7 Of the variables $(y_i^*, y_i, \boldsymbol{x}_i)$ only the pair $(y_i, \boldsymbol{x}_i)$ are observed. In this case, we say that y_i^* is a *latent* variable. Suppose

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= \mathbf{0} \\ y_i &= y_i^* + u_i \end{aligned}$$

where u_i is a measurement error satisfying

$$\begin{aligned} \mathbb{E}(\boldsymbol{x}_i u_i) &= \mathbf{0} \\ \mathbb{E}(y_i^* u_i) &= 0 \end{aligned}$$

Let $\widehat{\boldsymbol{\beta}}$ denote the OLS coefficient from the regression of y_i on $\boldsymbol{x}_i$.

(a) Is $\boldsymbol{\beta}$ the coefficient from the linear projection of y_i on $\boldsymbol{x}_i$?

(b) Is $\widehat{\boldsymbol{\beta}}$ consistent for $\boldsymbol{\beta}$ as $n \to \infty$?

(c) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ as $n \to \infty$.

Exercise 7.8 Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\sigma}^2 - \sigma^2\right)$ as $n \to \infty$.

Exercise 7.9 The model is

$$\begin{aligned} y_i &= x_i\beta + e_i \\ \mathbb{E}(e_i \mid x_i) &= 0 \end{aligned}$$

where $x_i \in \mathbb{R}$. Consider the two estimators

$$\begin{aligned} \widehat{\beta} &= \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \\ \widetilde{\beta} &= \frac{1}{n}\sum_{i=1}^n \frac{y_i}{x_i}. \end{aligned}$$

(a) Under the stated assumptions, are both estimators consistent for β?

(b) Are there conditions under which either estimator is efficient?

Exercise 7.10 In the homoskedastic regression model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with $\mathbb{E}(e_i \mid \boldsymbol{x}_i) = 0$ and $\mathbb{E}(e_i^2 \mid \boldsymbol{x}_i) = \sigma^2$, suppose $\widehat{\boldsymbol{\beta}}$ is the OLS estimate of $\boldsymbol{\beta}$ with covariance matrix estimate $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, based on a sample of size n. Let $\widehat{\sigma}^2$ be the estimate of σ^2. You wish to forecast an out-of-sample value of y_{n+1} given that $\boldsymbol{x}_{n+1} = \boldsymbol{x}$. Thus the available information is the sample $(\boldsymbol{y}, \boldsymbol{X})$, the estimates $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}, \widehat{\sigma}^2)$, the residuals $\widehat{\boldsymbol{e}}$, and the out-of-sample value of the regressors, $\boldsymbol{x}_{n+1}$.

(a) Find a point forecast of y_{n+1}.

(b) Find an estimate of the variance of this forecast.

Exercise 7.11 Take a regression model with i.i.d. observations (y_i, x_i) and scalar x_i

$$y_i = x_i\beta + e_i$$

$$\mathbb{E}(e_i \mid x_i) = 0$$

$$\Omega = \mathbb{E}\left(x_i^2 e_i^2\right)$$

Let $\widehat{\beta}$ be the OLS estimate of β with residuals $\widehat{e}_i = y_i - x_i\widehat{\beta}$. Consider the estimates of Ω

$$\widetilde{\Omega} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 e_i^2$$
$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \widehat{e}_i^2$$

(a) Find the asymptotic distribution of $\sqrt{n}\left(\widetilde{\Omega} - \Omega\right)$ as $n \to \infty$.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\Omega} - \Omega\right)$ as $n \to \infty$.

(c) How do you use the regression assumption $\mathbb{E}(e_i \mid x_i) = 0$ in your answer to (b)?

Exercise 7.12 Consider the model

$$\begin{aligned} y_i &= \alpha + \beta x_i + e_i \\ \mathbb{E}\left(e_i\right) &= 0 \\ \mathbb{E}\left(x_i e_i\right) &= 0 \end{aligned}$$

with both y_i and x_i scalar. Assuming $\alpha > 0$ and $\beta < 0$, suppose the parameter of interest is the area under the regression curve (e.g. consumer surplus), which is $A = -\alpha^2/2\beta$.

Let $\widehat{\boldsymbol{\theta}} = (\widehat{\alpha}, \widehat{\beta})'$ be the least-squares estimates of $\boldsymbol{\theta} = (\alpha, \beta)'$ so that $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to_d N(0, \boldsymbol{V}_{\boldsymbol{\theta}})$ and let $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ be a standard consistent estimate for $\boldsymbol{V}_{\boldsymbol{\theta}}$.

(a) Given the above, describe an estimator of A.

(b) Construct an asymptotic $(1-\eta)$ confidence interval for A.

Exercise 7.13 Consider an iid sample $\{y_i, x_i\}$ $i = 1, ..., n$ where y_i and x_i are scalar. Consider the reverse projection model

$$\begin{aligned} x_i &= y_i\gamma + u_i \\ \mathbb{E}\left(y_i u_i\right) &= 0 \end{aligned}$$

and define the parameter of interest as $\theta = 1/\gamma$

(a) Propose an estimator $\widehat{\gamma}$ of γ.

(b) Propose an estimator $\widehat{\theta}$ of θ.

(c) Find the asymptotic distribution of $\widehat{\theta}$.

(d) Find an asymptotic standard error for $\widehat{\theta}$.

Exercise 7.14 Take the model

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}\left(x_i e_i\right) &= 0 \end{aligned}$$

with both $\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}$, and define the parameter

$$\theta = \beta_1\beta_2$$

(a) What is the appropriate estimator $\widehat{\theta}$ for θ?

(b) Find the asymptotic distribution of $\widehat{\theta}$ under standard regularity conditions.

(c) Show how to calculate an asymptotic 95% confidence interval for θ.

Exercise 7.15 Take the linear model

$$\begin{aligned} y_i &= x_i\beta + e_i \\ \mathbb{E}\left(e_i \mid x_i\right) &= 0 \end{aligned}$$

with n observations and x_i is scalar (real-valued). Consider the estimator

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i^3 y_i}{\sum_{i=1}^n x_i^4}$$

Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ as $n \to \infty$.

Exercise 7.16 Out of an iid sample $(y_i, \boldsymbol{x}_i)$ of size n, you randomly take half the observations and estimate the least-squares regression of y_i on $\boldsymbol{x}_i$ using only this sub-sample.

$$y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{e}_i$$

Is the estimated slope coefficient $\widehat{\boldsymbol{\beta}}$ consistent for the population projection coefficient? Explain your reasoning.

Exercise 7.17 An economist reports a set of parameter estimates, including the coefficient estimates $\widehat{\beta}_1 = 1.0$, $\widehat{\beta}_2 = 0.8$, and standard errors $s(\widehat{\beta}_1) = 0.07$ and $s(\widehat{\beta}_2) = 0.07$. The author writes "The estimates show that β_1 is larger than β_2."

(a) Write down the formula for an asymptotic 95% confidence interval for $\theta = \beta_1 - \beta_2$, expressed as a function of $\widehat{\beta}_1$, $\widehat{\beta}_2$, $s(\widehat{\beta}_1)$, $s(\widehat{\beta}_2)$ and $\widehat{\rho}$, where $\widehat{\rho}$ is the estimated correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

(b) Can $\widehat{\rho}$ be calculated from the reported information?

(c) Is the author correct? Does the reported information support the author's claim?

Exercise 7.18 Suppose an economic model suggests

$$g(x) = \mathbb{E}\left(y_i \mid x_i = x\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$

where $x_i \in \mathbb{R}$. You have a random sample (y_i, x_i), $i = 1, ..., n$.

(a) Describe how to estimate $g(x)$ at a given value x.

(b) Describe (be specific) an appropriate confidence interval for $g(x)$.

Exercise 7.19 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned}$$

and suppose you have observations $i = 1, ..., 2n$. (The number of observations is $2n$.) You randomly split the sample in half, (each has n observations), calculate $\widehat{\boldsymbol{\beta}}_1$ by least-squares on the first sample, and $\widehat{\boldsymbol{\beta}}_2$ by least-squares on the second sample. What is the asymptotic distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\right)$?

Exercise 7.20 The data $\{y_i, \boldsymbol{x}_i, w_i\}$ is from a random sample, $i = 1, ..., n$. The parameter β is estimated by minimizing the criterion function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

That is $\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$.

(a) Find an explicit expression for $\widehat{\boldsymbol{\beta}}$.

(b) What population parameter $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}}$ estimating? (Be explicit about any assumptions you need to impose. But don't make more assumptions than necessary.)

(c) Find the probability limit for $\widehat{\boldsymbol{\beta}}$ as $n \to \infty$.

(d) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ as $n \to \infty$.

Exercise 7.21 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \\ \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) &= \sigma_i^2 = \boldsymbol{z}_i'\boldsymbol{\gamma} \end{aligned}$$

where $\boldsymbol{z}_i$ is a (vector) function of $\boldsymbol{x}_i$. The sample is $i = 1, ..., n$ with iid observations. For simplicity, assume that $\boldsymbol{z}_i'\boldsymbol{\gamma} > 0$ for all $\boldsymbol{z}_i$. Suppose you are interested in forecasting y_{n+1} given $\boldsymbol{x}_{n+1} = \boldsymbol{x}$ and $\boldsymbol{z}_{n+1} = \boldsymbol{z}$ for some out-of-sample observation $n+1$. Describe how you would construct a point forecast and a forecast interval for y_{n+1}.

Exercise 7.22 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \\ z_i &= \left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\gamma + u_i \\ \mathbb{E}\left(u_i \mid \boldsymbol{x}_i\right) &= 0 \end{aligned}$$

Your goal is to estimate γ. (Note that γ is scalar.) You use a two-step estimator:

- Estimate $\widehat{\boldsymbol{\beta}}$ by least-squares of y_i on $\boldsymbol{x}_i$.
- Estimate $\widehat{\gamma}$ by least-squares of z_i on $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$.

(a) Show that $\widehat{\gamma}$ is consistent for γ.

(b) Find the asymptotic distribution of $\widehat{\gamma}$ when $\gamma = 0$.

Exercise 7.23 The model is

$$\begin{aligned} y_i &= x_i\beta + e_i \\ \mathbb{E}\left(e_i \mid x_i\right) &= 0 \end{aligned}$$

where $x_i \in R$. Consider the the estimator

$$\widetilde{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{y_i}{x_i}.$$

Find conditions under which $\widetilde{\beta}$ is consistent for β as $n \to \infty$.

Exercise 7.24 Of the random variables $(y_i^*, y_i, \boldsymbol{x}_i)$ only the pair $(y_i, \boldsymbol{x}_i)$ are observed. (In this case, we say that y_i^* is a *latent* variable.) Suppose $\mathbb{E}\left(y_i^* \mid \boldsymbol{x}_i\right) = \boldsymbol{x}_i'\boldsymbol{\beta}$ and $y = y_i^* + u_i$, where u_i is a measurement error satisfying $\mathbb{E}\left(u_i \mid y_i^*, \boldsymbol{x}_i\right) = 0$. Let $\widehat{\boldsymbol{\beta}}$ denote the OLS coefficient from the regression of y_i on $\boldsymbol{x}_i$.

(a) Find $\mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right)$.

(b) Is $\widehat{\boldsymbol{\beta}}$ consistent for $\boldsymbol{\beta}$ as $n \to \infty$?

(c) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ as $n \to \infty$.

Exercise 7.25 The parameter of β is defined in the model

$$y_i = x_i^*\beta + e_i$$

where e_i is independent of x_i^*, $\mathbb{E}\left(e_i\right) = 0$, $\mathbb{E}\left(e_i^2\right) = \sigma^2$. The observables are (y_i, x_i) where

$$x_i = x_i^* v_i$$

and $v_i > 0$ is random measurement error. Assume that v_i is independent of x_i^* and e_i. Also assume that x_i and x_i^* are non-negative and real-valued. Consider the least-squares estimator $\widehat{\beta}$ for β.

(a) Find the plim of $\widehat{\beta}$, expressed in terms of β and moments of (x_i, v_i, e_i)

(b) Can you find a non-trivial condition under which $\widehat{\beta}$ is consisent for β? (By non-trivial, we mean something other than $v_i = 1$.)

Exercise 7.26 Take the standard model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned}$$

For a positive function $w(\boldsymbol{x})$, let $w_i = w(\boldsymbol{x}_i)$. Consider the estimator

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^n w_i \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^n w_i \boldsymbol{x}_i y_i\right).$$

Find the probability limit (as $n \to \infty$) of $\widetilde{\boldsymbol{\beta}}$. (Do you need to add an assumption?) Is $\widetilde{\boldsymbol{\beta}}$ consistent for $\widetilde{\boldsymbol{\beta}}$? If not, under what assumption is $\widetilde{\boldsymbol{\beta}}$ consistent for $\boldsymbol{\beta}$?

Exercise 7.27 Take the regression model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \\ \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) &= \sigma_i^2 \end{aligned}$$

with $\boldsymbol{x}_i \in R^k$. Assume that $\Pr\left(e_i = 0\right) = 0$. Consider the infeasible estimator

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^n e_i^{-2} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^n e_i^{-2} \boldsymbol{x}_i y_i\right).$$

This is a WLS estimator using the weights e_i^{-2}.

(a) Find the asymptotic distribution of $\widetilde{\boldsymbol{\beta}}$

(b) Contrast your result with the asymptotic distribution of infeasible GLS.

Exercise 7.28 The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0. \end{aligned}$$

An econometrician is worried about the impact of some unusually large values of the regressors. The model is thus estimated on the subsample for which $|\boldsymbol{x}_i| \leq c$, for some fixed c. Let $\widetilde{\boldsymbol{\beta}}$ denote the OLS estimator on this subsample. It equals

$$\widetilde{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' 1\left(|\boldsymbol{x}_i| \leq c\right)\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i 1\left(|\boldsymbol{x}_i| \leq c\right)\right)$$

where $1(\cdot)$ denotes the indicator function.

(a) Show that $\widetilde{\boldsymbol{\beta}} \to_p \boldsymbol{\beta}$.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$

Exercise 7.29 As in Exercise 3.24, use the CPS dataset and the subsample of white male Hispanics. Estimate the regression

$$\widehat{\log(Wage)} = \beta_1\, education + \beta_2\, experience + \beta_3\, experience^2/100 + \beta_4.$$

(a) Report the coefficients and robust standard errors.

(b) Let θ be the ratio of the return to one year of education to the return to one year of experience. Write θ as a function of the regression coefficients and variables. Compute $\widehat{\theta}$ from the estimated model.

(c) Write out the formula for the asymptotic standard error for $\widehat{\theta}$ as a function of the covariance matrix for $\widehat{\boldsymbol{\beta}}$. Compute $\widehat{s}(\widehat{\theta})$ from the estimated model.

(d) Construct a 90% asymptotic confidence interval for θ from the estimated model.

(e) Compute the regression function at $edu = 12$ and experience=20. Compute a 95% confidence interval for the regression function at this point.

(f) Consider an out-of-sample individual with 16 years of education and 5 years experience. Construct an 80% forecast interval for their log wage and wage. [To obtain the forecast interval for the wage, apply the exponential function to both endpoints.]

Chapter 8

Restricted Estimation

8.1 Introduction

In the linear projection model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned}$$

a common task is to impose a constraint on the coefficient vector $\boldsymbol{\beta}$. For example, partitioning $\boldsymbol{x}_i' = (\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}')$ and $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)$, a typical constraint is an exclusion restriction of the form $\boldsymbol{\beta}_2 = \mathbf{0}$. In this case the constrained model is

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned}$$

At first glance this appears the same as the linear projection model, but there is one important difference: the error e_i is uncorrelated with the entire regressor vector $\boldsymbol{x}_i' = (\boldsymbol{x}_{1i}', \boldsymbol{x}_{2i}')$ not just the included regressor $\boldsymbol{x}_{1i}$.

In general, a set of q linear constraints on $\boldsymbol{\beta}$ takes the form

$$\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c} \tag{8.1}$$

where $\boldsymbol{R}$ is $k \times q$, $\operatorname{rank}(\boldsymbol{R}) = q < k$ and $\boldsymbol{c}$ is $q \times 1$. The assumption that $\boldsymbol{R}$ is full rank means that the constraints are linearly independent (there are no redundant or contradictory constraints). We can define the restricted parameter space $\boldsymbol{B}_{\boldsymbol{R}}$ as the set of values of $\boldsymbol{\beta}$ which satisfy (8.1), that is

$$\boldsymbol{B}_{\boldsymbol{R}} = \left\{\boldsymbol{\beta} : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}\right\}$$

The constraint $\boldsymbol{\beta}_2 = \mathbf{0}$ discussed above is a special case of the constraint (8.1) with

$$\boldsymbol{R} = \begin{pmatrix} \mathbf{0} \\ \boldsymbol{I} \end{pmatrix}, \tag{8.2}$$

a selector matrix, and $\boldsymbol{c} = \mathbf{0}$.

Another common restriction is that a set of coefficients sum to a known constant, i.e. $\beta_1 + \beta_2 = 1$. This constraint arises in a constant-return-to-scale production function. Other common restrictions include the equality of coefficients $\beta_1 = \beta_2$, and equal and offsetting coefficients $\beta_1 = -\beta_2$.

A typical reason to impose a constraint is that we believe (or have information) that the constraint is true. By imposing the constraint we hope to improve estimation efficiency. The goal is to obtain consistent estimates with reduced variance relative to the unconstrained estimator.

The questions then arise: How should we estimate the coefficient vector $\boldsymbol{\beta}$ imposing the linear restriction (8.1)? If we impose such constraints, what is the sampling distribution of the resulting estimator? How should we calculate standard errors? These are the questions explored in this chapter.

8.2 Constrained Least Squares

An intuitively appealing method to estimate a constrained linear projection is to minimize the least-squares criterion subject to the constraint $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}$.

The constrained least-squares estimator is

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}} SSE(\boldsymbol{\beta}) \tag{8.3}$$

where

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}. \tag{8.4}$$

The estimator $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ minimizes the sum of squared errors over all $\boldsymbol{\beta}$ such that $\boldsymbol{\beta} \in \boldsymbol{B}_{\boldsymbol{R}}$, or equivalently such that the restriction (8.1) holds. We call $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ the **constrained least-squares** (CLS) estimator. We follow the convention of using a tilde "˜" rather than a hat "ˆ" to indicate that $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ is a restricted estimator in contrast to the unrestricted least-squares estimator $\widehat{\boldsymbol{\beta}}$, and write it as $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ to be clear that the estimation method is CLS.

One method to find the solution to (8.3) uses the technique of Lagrange multipliers. The problem (8.3) is equivalent to the minimization of the Lagrangian

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} SSE(\boldsymbol{\beta}) + \boldsymbol{\lambda}'\left(\boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{c}\right) \tag{8.5}$$

over $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is an $s \times 1$ vector of Lagrange multipliers. The first-order conditions for minimization of (8.5) are

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}(\widetilde{\boldsymbol{\beta}}_{\text{cls}}, \widetilde{\boldsymbol{\lambda}}_{\text{cls}}) = -\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{X}'\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{\text{cls}} + \boldsymbol{R}\widetilde{\boldsymbol{\lambda}}_{\text{cls}} = \boldsymbol{0} \tag{8.6}$$

and

$$\frac{\partial}{\partial \boldsymbol{\lambda}} \mathcal{L}(\widetilde{\boldsymbol{\beta}}_{\text{cls}}, \widetilde{\boldsymbol{\lambda}}_{\text{cls}}) = \boldsymbol{R}'\widetilde{\boldsymbol{\beta}} - \boldsymbol{c} = \boldsymbol{0}. \tag{8.7}$$

Premultiplying (8.6) by $\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ we obtain

$$-\boldsymbol{R}'\widehat{\boldsymbol{\beta}} + \boldsymbol{R}'\widetilde{\boldsymbol{\beta}}_{\text{cls}} + \boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\widetilde{\boldsymbol{\lambda}}_{\text{cls}} = \boldsymbol{0} \tag{8.8}$$

where $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$ is the unrestricted least-squares estimator. Imposing $\boldsymbol{R}'\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{c} = \boldsymbol{0}$ from (8.7) and solving for $\widetilde{\boldsymbol{\lambda}}_{\text{cls}}$ we find

$$\widetilde{\boldsymbol{\lambda}}_{\text{cls}} = \left[\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right).$$

Notice that $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} > 0$ and $\boldsymbol{R}$ full rank imply that $\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R} > 0$ and is hence invertible. (See Section A.9.)

Substituting this expression into (8.6) and solving for $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ we find the solution to the constrained minimization problem (8.3)

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left[\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right). \tag{8.9}$$

(See Exercise 8.5 to verify that (8.9) satisfies (8.1).)

This is a general formula for the CLS estimator. It also can be written as

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right). \tag{8.10}$$

The CLS residuals are

$$\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}_{\text{cls}}$$

and the $n \times 1$ vector of residuals are written in vector notation as $\widetilde{\boldsymbol{e}}$.

In Stata, constrainded least squares is implemented using the `cnsreg` command.

8.3 Exclusion Restriction

While (8.9) is a general formula for the CLS estimator, in most cases the estimator can be found by applying least-squares to a reparameterized equation. To illustrate, let us return to the first example presented at the beginning of the chapter – a simple exclusion restriction. Recall the unconstrained model is

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \tag{8.11}$$

the exclusion restriction is $\boldsymbol{\beta}_2 = \mathbf{0}$, and the constrained equation is

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_i. \tag{8.12}$$

In this setting the CLS estimator is OLS of y_i on x_{1i}. (See Exercise 8.1.) We can write this as

$$\widetilde{\boldsymbol{\beta}}_1 = \left(\sum_{i=1}^{n} \boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_{1i}y_i\right). \tag{8.13}$$

The CLS estimator of the entire vector $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)$ is

$$\widetilde{\boldsymbol{\beta}} = \begin{pmatrix} \widetilde{\boldsymbol{\beta}}_1 \\ \mathbf{0} \end{pmatrix}. \tag{8.14}$$

It is not immediately obvious, but (8.9) and (8.14) are algebraically (and numerically) equivalent. To see this, the first component of (8.9) with (8.2) is

$$\widetilde{\boldsymbol{\beta}}_1 = \begin{pmatrix} \boldsymbol{I} & \mathbf{0} \end{pmatrix}\left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\begin{pmatrix} \mathbf{0} \\ \boldsymbol{I} \end{pmatrix}\left[\begin{pmatrix} \mathbf{0} & \boldsymbol{I} \end{pmatrix}\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\begin{pmatrix} \mathbf{0} \\ \boldsymbol{I} \end{pmatrix}\right]^{-1}\begin{pmatrix} \mathbf{0} & \boldsymbol{I} \end{pmatrix}\widehat{\boldsymbol{\beta}}\right].$$

Using (3.39) this equals

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_1 &= \widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{Q}}^{12}\left(\widehat{\boldsymbol{Q}}^{22}\right)^{-1}\widehat{\boldsymbol{\beta}}_2 \\
&= \widehat{\boldsymbol{\beta}}_1 + \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{22\cdot 1}\widehat{\boldsymbol{\beta}}_2 \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{1y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{2y}\right) \\
&+ \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{22\cdot 1}\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\left(\widehat{\boldsymbol{Q}}_{2y} - \widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}\right) \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{1y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}\right) \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}\right)\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y} \\
&= \widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}
\end{aligned}$$

which is (8.14) as originally claimed.

8.4 Finite Sample Properties

In this section we explore some of the properties of the CLS estimator in the linear regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{8.15}$$

$$\mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) = 0. \tag{8.16}$$

First, it is useful to write the estimator, and the residuals, as linear functions of the error vector. These are algebraic relationships and do not rely on the linear regression assumptions.

Theorem 8.4.1 *Define* $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ *and*

$$\boldsymbol{A} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Then

1. $\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c} = \boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}$
2. $\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta} = \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' - \boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}$
3. $\widetilde{\boldsymbol{e}} = \left(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}$
4. $\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}$ *is symmetric and idempotent*
5. $\operatorname{tr}\left(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}\right) = n - k + q.$

See Exercise 8.6.

Given the linearity of Theorem 8.4.1.2, it is not hard to show that the CLS estimator is unbiased for $\boldsymbol{\beta}$

Theorem 8.4.2 *In the linear regression model (8.15-(8.16) under 8.6.1,* $\mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$.

See Exercise 8.7.

Given the linearity we can also calculate the variance matrix of $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$. For this we will add the assumption of conditional homoskedasticity to simplify the expression.

Theorem 8.4.3 *In the homoskedastic linear regression model (8.15-(8.16) with* $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$*, under 8.6.1,*

$$\begin{aligned}\boldsymbol{V}_{\widetilde{\boldsymbol{\beta}}}^0 &= \operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}} \mid \boldsymbol{X}\right) \\ &= \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\sigma^2\end{aligned}$$

See Exercise 8.8. We use the $\boldsymbol{V}_{\widetilde{\boldsymbol{\beta}}}^0$ notation to emphasize that this is the variance matrix under the assumption of conditional homoskedasticity.

For inference we need an estimate of $\boldsymbol{V}_{\widetilde{\boldsymbol{\beta}}}^0$. A natural estimator is

$$\widehat{\boldsymbol{V}}_{\widetilde{\boldsymbol{\beta}}}^0 = \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)s_{\text{cls}}^2$$

where

$$s_{\text{cls}}^2 = \frac{1}{n-k+q}\sum_{i=1}^{n}\widetilde{e}_i^2 \tag{8.17}$$

is a biased-corrected estimator of σ^2. Standard errors for the components of $\boldsymbol{\beta}$ are then found by taking the squares roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widetilde{\boldsymbol{\beta}}}$, for example

$$s(\widehat{\beta}_j) = \sqrt{\left[\widehat{\boldsymbol{V}}^0_{\widetilde{\boldsymbol{\beta}}}\right]_{jj}}.$$

The estimator (8.17) has the property that it is unbiased for σ^2 under conditional homoskedasticity. To see this, using the properties of Theorem 8.4.1,

$$\begin{aligned}(n-k+q)\, s^2_{\text{cls}} &= \widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}} \\ &= \boldsymbol{e}'\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e} \\ &= \boldsymbol{e}'\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}. \end{aligned} \tag{8.18}$$

We defer the remainder of the proof to Exercise 8.9.

Theorem 8.4.4 *In the homoskedastic linear regression model (8.15-(8.16) with* $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$*, under 8.6.1,* $\mathbb{E}\left(s^2_{\text{cls}} \mid \boldsymbol{X}\right) = \sigma^2$ *and* $\mathbb{E}\left(\widehat{\boldsymbol{V}}^0_{\widetilde{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) = \boldsymbol{V}^0_{\widetilde{\boldsymbol{\beta}}}$.

Now consider the distributional properties in the normal regression model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ e_i &\sim \mathrm{N}(0, \sigma^2). \end{aligned}$$

By the linearity of Theorem 8.4.1.2, conditional on $\boldsymbol{X}$, $\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta}$ is normal. Given Theorems 8.4.2 and 8.4.3, we deduce that $\widetilde{\boldsymbol{\beta}}_{\text{cls}} \sim \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{V}^0_{\widetilde{\boldsymbol{\beta}}})$.

Similarly, from Exericise 8.4.1 we know $\widetilde{\boldsymbol{e}} = \left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}$ is linear in $\boldsymbol{e}$ so is also conditionally normal. Furthermore, since $\left(\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}-\boldsymbol{X}\boldsymbol{A}\right) = 0$, $\widetilde{\boldsymbol{e}}$ and $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ are uncorrelated and thus independent. Thus s^2_{cls} and $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ are independent.

From (8.18) and the fact that $\boldsymbol{I}-\boldsymbol{P}+\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'$ is idempotent with rank $n-k+q$, it follows that

$$s^2_{\text{cls}} \sim \sigma^2 \chi^2_{n-k+q} / \left(n-k+q\right).$$

It follows that the t-statistic has the exact distribution

$$\begin{aligned} T &= \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)} \\ &\sim \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-k+q} \Big/ \left(n-k+q\right)}} \\ &\sim t_{n-k+q} \end{aligned}$$

a student t distribution with $n-k+q$ degrees of freedom.

The relevance of this calculation is that the "degrees of freedom" for a CLS regression problem equal $n-k+q$ rather than $n-k$ as in the OLS regression problem. Essentially, the model has $k-q$ free parameters instead of k. Another way of thinking about this is that estimation of a model with k coefficients and q restrictions is equivalent to estimation with $k-q$ coefficients.

We summarize the properties of the normal regression model

Theorem 8.4.5 *In the normal linear regression model linear regression model (8.15-(8.16), under 8.6.1,*

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} \sim \mathrm{N}(\boldsymbol{\beta}, \boldsymbol{V}^0_{\widetilde{\boldsymbol{\beta}}})$$
$$\frac{(n-k+q)\, s^2_{\text{cls}}}{\sigma^2} \sim \chi^2_{n-k+q}$$
$$T \sim t_{n-k+q}$$

An interesting relationship is that in the homoskedastic regression model

$$\begin{aligned}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}, \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) &= \mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta}\right)'\right) \\ &= \mathbb{E}\left(\left(\boldsymbol{A}\boldsymbol{X}'\right)\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \boldsymbol{X}\boldsymbol{A}\right)\right)\sigma^2 = 0\end{aligned}$$

so $\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}$ and $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ are uncorrelated and hence independent. One corollary is

$$\operatorname{cov}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}, \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) = \operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)$$

A second corollary is

$$\begin{aligned}\operatorname{var}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) &= \operatorname{var}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}\right) - \operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2.\end{aligned} \tag{8.19}$$

This also shows us the difference between the CLS and OLS variances

$$\operatorname{var}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}\right) - \operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \geq 0$$

the final equality meaning positive semi-definite. It follows that $\operatorname{var}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}\right) \geq \operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)$ in the positive definite sense, and thus CLS is more efficient than OLS. Both estimators are unbiased (in the linear regression model), and CLS has a lower variance matrix (in the linear homoskedastic regression model).

The relationship (8.19) is rather interesting and will appear again. The expression says that the variance of the difference between the estimators is equal to the difference between the variances. This is rather special. It occurs (generically) when we are comparing an efficient and an inefficient estimator. We call (8.19) the **Hausmann Equality** as it was first pointed out in econometrics by Hausman (1978).

8.5 Minimum Distance

The previous section explored the finite sample distribution theory under the assumptions of the linear regression model, homoskedastic regression model, and normal regression model. We now return to the general projection model where we do not impose linearity, homoskedasticity, nor normality. We are interested in the question: Can we do better than CLS in this setting?

A minimum distance estimator tries to find a parameter value which satisfies the constraint which is as close as possible to the unconstrained estimate. Let $\widehat{\boldsymbol{\beta}}$ be the unconstrained least-squares estimator, and for some $k \times k$ positive definite weight matrix $\widehat{\boldsymbol{W}} > 0$ define the quadratic criterion function

$$J(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\widehat{\boldsymbol{W}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{8.20}$$

This is a (squared) weighted Euclidean distance between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. $J(\boldsymbol{\beta})$ is small if $\boldsymbol{\beta}$ is close to $\widehat{\boldsymbol{\beta}}$, and is minimized at zero only if $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. A **minimum distance estimator** $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ for $\boldsymbol{\beta}$ minimizes $J(\boldsymbol{\beta})$ subject to the constraint (8.1), that is,

$$\widetilde{\boldsymbol{\beta}}_{\text{md}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}}\ J(\boldsymbol{\beta}). \tag{8.21}$$

The CLS estimator is the special case when $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{xx}$, and we write this criterion function as

$$J^0(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{8.22}$$

To see the equality of CLS and minimum distance, rewrite the least-squares criterion as follows. Write the unconstrained least-squares fitted equation as $y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{e}_i$ and substitute this equation into $SSE(\boldsymbol{\beta})$ to obtain

$$\begin{aligned} SSE(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\ &= \sum_{i=1}^{n} \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{e}_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\ &= \sum_{i=1}^{n} \widehat{e}_i^2 + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &= n\widehat{\sigma}^2 + J^0(\boldsymbol{\beta}) \end{aligned} \tag{8.23}$$

where the third equality uses the fact that $\sum_{i=1}^n \boldsymbol{x}_i\widehat{e}_i = 0$, and the last line uses $\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' = n\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}$. The expression (8.23) only depends on $\boldsymbol{\beta}$ through $J^0(\boldsymbol{\beta})$. Thus minimization of $SSE(\boldsymbol{\beta})$ and $J^0(\boldsymbol{\beta})$ are equivalent, and hence $\widetilde{\boldsymbol{\beta}}_{\text{md}} = \widetilde{\boldsymbol{\beta}}_{\text{cls}}$ when $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}xx}$.

We can solve for $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ explicitly by the method of Lagrange multipliers. The Lagrangian is

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} J\left(\boldsymbol{\beta}, \widehat{\boldsymbol{W}}\right) + \boldsymbol{\lambda}'\left(\boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{c}\right)$$

which is minimized over $(\boldsymbol{\beta}, \boldsymbol{\lambda})$. The solution is

$$\widetilde{\boldsymbol{\lambda}}_{\text{md}} = n\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right) \tag{8.24}$$

$$\widetilde{\boldsymbol{\beta}}_{\text{md}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right). \tag{8.25}$$

(See Exercise 8.10.) Comparing (8.25) with (8.10) we can see that $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ specializes to $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ when we set $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}$.

An obvious question is which weight matrix $\widehat{\boldsymbol{W}}$ is best. We will address this question after we derive the asymptotic distribution for a general weight matrix.

8.6 Asymptotic Distribution

We first show that the class of minimum distance estimators are consistent for the population parameters when the constraints are valid.

Assumption 8.6.1 $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}$ *where* $\boldsymbol{R}$ *is* $k \times q$ *with* $\operatorname{rank}(\boldsymbol{R}) = q$.

Assumption 8.6.2 $\widehat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W} > 0$.

Theorem 8.6.1 ***Consistency***
Under Assumptions 7.1.1, 8.6.1, and 8.6.2, $\widetilde{\boldsymbol{\beta}}_{\text{md}} \xrightarrow{p} \boldsymbol{\beta}$ *as* $n \to \infty$.

For a proof, see Exercise 8.11.

Theorem 8.6.1 shows that consistency holds for any weight matrix with a positive definite limit, so the result includes the CLS estimator.

Similarly, the constrained estimators are asymptotically normally distributed.

Theorem 8.6.2 ***Asymptotic Normality***
Under Assumptions 7.1.2, 8.6.1, and 8.6.2,

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\text{md}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right) \tag{8.26}$$

as $n \to \infty$, *where*

$$\begin{aligned} \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) = \boldsymbol{V}_{\boldsymbol{\beta}} &- \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}} \\ &- \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}^{-1} \\ &+ \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}^{-1} \end{aligned} \tag{8.27}$$

and $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$

For a proof, see Exercise 8.12.

Theorem 8.6.2 shows that the minimum distance estimator is asymptotically normal for all positive definite weight matrices. The asymptotic variance depends on $\boldsymbol{W}$. The theorem includes the CLS estimator as a special case by setting $\boldsymbol{W} = \boldsymbol{Q}_{xx}$.

Theorem 8.6.3 ***Asymptotic Distribution of CLS Estimator***
Under Assumptions 7.1.2 and 8.6.1, as $n \to \infty$

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\text{cls}}\right)$$

where

$$\begin{aligned} \boldsymbol{V}_{\text{cls}} = \boldsymbol{V}_{\boldsymbol{\beta}} &- \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}} \\ &- \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1} \\ &+ \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}_{\boldsymbol{xx}}^{-1} \end{aligned}$$

For a proof, see Exercise 8.13.

8.7 Variance Estimation and Standard Errors

Earlier we intruduce the covariance matrix estimator under the assumption of conditional homoskedasticity. We now introduce an estimator which does not impose homoskedasticity.

The asymptotic covariance matrix $\boldsymbol{V}_{\text{cls}}$ may be estimated by replacing $\boldsymbol{V}_{\boldsymbol{\beta}}$ with a consistent estimates such as $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$. A more efficient estimate is obtained by using the restricted estimates. Given the constrained least-squares squares residuals $\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ we can estimate the matrix $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'e_i^2\right)$ by

$$\widetilde{\boldsymbol{\Omega}} = \frac{1}{n-k+q}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\widetilde{e}_i^2.$$

Notice that we have defined $\widetilde{\boldsymbol{\Omega}}$ using an adjusted degrees of freedom. This is an ad hoc adjustment designed to mimic that used for estimation of the error variance σ^2. Given $\widetilde{\boldsymbol{\Omega}}$ the moment estimator of $\boldsymbol{V}_{\boldsymbol{\beta}}$ is

$$\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}} = \widehat{\boldsymbol{Q}}_{xx}^{-1}\widetilde{\boldsymbol{\Omega}}\widehat{\boldsymbol{Q}}_{xx}^{-1}$$

and that for $\boldsymbol{V}_{\text{cls}}$ is

$$\begin{aligned}\widetilde{\boldsymbol{V}}_{\text{cls}} = \widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}} &- \widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}} \\ &- \widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{xx}^{-1} \\ &+ \widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{xx}^{-1}.\end{aligned}$$

We can calculate standard errors for any linear combination $\boldsymbol{h}'\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ so long as $\boldsymbol{h}$ does not lie in the range space of $\boldsymbol{R}$. A standard error for $\boldsymbol{h}'\widetilde{\boldsymbol{\beta}}$ is

$$s(\boldsymbol{h}'\widetilde{\boldsymbol{\beta}}_{\text{cls}}) = \left(n^{-1}\boldsymbol{h}'\widetilde{\boldsymbol{V}}_{\text{cls}}\boldsymbol{h}\right)^{1/2}.$$

8.8 Efficient Minimum Distance Estimator

Theorem 8.6.2 shows that the minimum distance estimators, which include CLS as a special case, are asymptotically normal with an asymptotic covariance matrix which depends on the weight matrix $\boldsymbol{W}$. The asymptotically optimal weight matrix is the one which minimizes the asymptotic variance $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})$. This turns out to be $\boldsymbol{W} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$ as is shown in Theorem 8.8.1 below. Since $\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$ is unknown this weight matrix cannot be used for a feasible estimator, but we can replace $\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$ with a consistent estimate $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}$ and the asymptotic distribution (and efficiency) are unchanged. We call the minimum distance estimator setting $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}$ the **efficient minimum distance estimator** and takes the form

$$\widetilde{\boldsymbol{\beta}}_{\text{emd}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right). \tag{8.28}$$

The asymptotic distribution of (8.28) can be deduced from Theorem 8.6.2. (See Exercises 8.14 and 8.15.)

Theorem 8.8.1 ***Efficient Minimum Distance Estimator***
Under Assumptions 7.1.2 and 8.6.1,

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\text{emd}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta},\text{emd}}\right)$$

as $n \to \infty$, *where*

$$\boldsymbol{V}_{\boldsymbol{\beta},\text{emd}} = \boldsymbol{V}_{\boldsymbol{\beta}} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}. \tag{8.29}$$

Since

$$\boldsymbol{V}_{\boldsymbol{\beta},\text{emd}} \leq \boldsymbol{V}_{\boldsymbol{\beta}} \tag{8.30}$$

the estimator (8.28) has lower asymptotic variance than the unrestricted estimator. Furthermore, for any $\boldsymbol{W}$,

$$\boldsymbol{V}_{\boldsymbol{\beta},\text{emd}} \leq \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) \tag{8.31}$$

so (8.28) is asymptotically efficient in the class of minimum distance estimators.

Theorem 8.8.1 shows that the minimum distance estimator with the smallest asymptotic variance is (8.28). One implication is that the constrained least squares estimator is generally inefficient. The interesting exception is the case of conditional homoskedasticity, in which case the optimal weight matrix is $\boldsymbol{W} = \left(\boldsymbol{V}^0_{\boldsymbol{\beta}}\right)^{-1}$ so in this case CLS is an efficient minimum distance estimator. Otherwise when the error is conditionally heteroskedastic, there are asymptotic efficiency gains by using minimum distance rather than least squares.

The fact that CLS is generally inefficient is counter-intuitive and requires some reflection to understand. Standard intuition suggests to apply the same estimation method (least squares) to the unconstrained and constrained models, and this is the most common empirical practice. But Theorem 8.8.1 shows that this is not the efficient estimation method. Instead, the efficient minimum distance estimator has a smaller asymptotic variance. Why? The reason is that the least-squares estimator does not make use of the regressor $\boldsymbol{x}_{2i}$. It ignores the information $\mathbb{E}\left(\boldsymbol{x}_{2i}e_i\right) = \mathbf{0}$. This information is relevant when the error is heteroskedastic and the excluded regressors are correlated with the included regressors.

Inequality (8.30) shows that the efficient minimum distance estimator $\widetilde{\boldsymbol{\beta}}_{\text{emd}}$ has a smaller asymptotic variance than the unrestricted least squares estimator $\widehat{\boldsymbol{\beta}}$. This means that estimation is more efficient by imposing correct restrictions when we use the minimum distance method.

8.9 Exclusion Restriction Revisited

We return to the example of estimation with a simple exclusion restriction. The model is

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i$$

with the exclusion restriction $\boldsymbol{\beta}_2 = \mathbf{0}$. We have introduced three estimators of $\boldsymbol{\beta}_1$. The first is unconstrained least-squares applied to (8.11), which can be written as

$$\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{1y\cdot 2}.$$

From Theorem 7.33 and equation (7.20) its asymptotic variance is

$$\operatorname{avar}(\widehat{\boldsymbol{\beta}}_1) = \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\boldsymbol{\Omega}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{\Omega}_{21} - \boldsymbol{\Omega}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{\Omega}_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot 2}^{-1}.$$

The second estimator of $\boldsymbol{\beta}_1$ is the CLS estimator, which can be written as

$$\widetilde{\boldsymbol{\beta}}_{1,\text{cls}} = \widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}.$$

Its asymptotic variance can be deduced from Theorem 8.6.3, but it is simpler to apply the CLT directly to show that

$$\text{avar}(\widetilde{\boldsymbol{\beta}}_{1,\text{cls}}) = \boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{11}\boldsymbol{Q}_{11}^{-1}. \tag{8.32}$$

The third estimator of $\boldsymbol{\beta}_1$ is the efficient minimum distance estimator. Applying (8.28), it equals

$$\widetilde{\boldsymbol{\beta}}_{1,\text{md}} = \widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{V}}_{12}\widehat{\boldsymbol{V}}_{22}^{-1}\widehat{\boldsymbol{\beta}}_2 \tag{8.33}$$

where we have partitioned

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\boldsymbol{V}}_{11} & \widehat{\boldsymbol{V}}_{12} \\ \widehat{\boldsymbol{V}}_{21} & \widehat{\boldsymbol{V}}_{22} \end{bmatrix}.$$

From Theorem 8.8.1 its asymptotic variance is

$$\text{avar}(\widetilde{\boldsymbol{\beta}}_{1,\text{md}}) = \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}. \tag{8.34}$$

See Exercise 8.16 to verify equations (8.32), (8.33), and (8.34).

In general, the three estimators are different, and they have different asymptotic variances.

It is quite instructive to compare the asymptotic variances of the CLS and unconstrained least-squares estimators to assess whether or not the constrained estimator is necessarily more efficient than the unconstrained estimator.

First, consider the case of conditional homoskedasticity. In this case the two covariance matrices simplify to

$$\text{avar}(\widehat{\boldsymbol{\beta}}_1) = \sigma^2\boldsymbol{Q}_{11\cdot 2}^{-1}$$

and

$$\text{avar}(\widetilde{\boldsymbol{\beta}}_{1,\text{cls}}) = \sigma^2\boldsymbol{Q}_{11}^{-1}.$$

If $\boldsymbol{Q}_{12} = 0$ (so $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are orthogonal) then these two variance matrices are equal and the two estimators have equal asymptotic efficiency. Otherwise, since $\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} \geq 0$, then $\boldsymbol{Q}_{11} \geq \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$, and consequently

$$\boldsymbol{Q}_{11}^{-1}\sigma^2 \leq \left(\boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)^{-1}\sigma^2.$$

This means that under conditional homoskedasticity, $\widetilde{\boldsymbol{\beta}}_{1,\text{cls}}$ has a lower asymptotic variance matrix than $\widehat{\boldsymbol{\beta}}_1$. Therefore in this context, constrained least-squares is more efficient than unconstrained least-squares. This is consistent with our intuition that imposing a correct restriction (excluding an irrelevant regressor) improves estimation efficiency.

However, in the general case of conditional heteroskedasticity this ranking is not guaranteed. In fact what is really amazing is that the variance ranking can be reversed. The CLS estimator can have a larger asymptotic variance than the unconstrained least squares estimator.

To see this let's use the simple heteroskedastic example from Section 7.4. In that example, $Q_{11} = Q_{22} = 1$, $Q_{12} = \dfrac{1}{2}$, $\Omega_{11} = \Omega_{22} = 1$, and $\Omega_{12} = \dfrac{7}{8}$. We can calculate (see Exercise 8.17) that $\boldsymbol{Q}_{11\cdot 2} = \dfrac{3}{4}$ and

$$\text{avar}(\widehat{\boldsymbol{\beta}}_1) = \frac{2}{3} \tag{8.35}$$

$$\text{avar}(\widetilde{\boldsymbol{\beta}}_{1,\text{cls}}) = 1 \tag{8.36}$$

$$\text{avar}(\widetilde{\boldsymbol{\beta}}_{1,\text{md}}) = \frac{5}{8}. \tag{8.37}$$

Thus the restricted least-squares estimator $\widetilde{\boldsymbol{\beta}}_{1,\text{cls}}$ has a larger variance than the unrestricted least-squares estimator $\widehat{\boldsymbol{\beta}}_1$! The minimum distance estimator has the smallest variance of the three, as expected.

What we have found is that when the estimation method is least-squares, deleting the irrelevant variable x_{2i} can actually increase estimation variance, or equivalently, adding an irrelevant variable can actually decrease the estimation variance.

To repeat this unexpected finding, we have shown in a very simple example that it is possible for least-squares applied to the short regression (8.12) to be less efficient for estimation of $\boldsymbol{\beta}_1$ than least-squares applied to the long regression (8.11), even though the constraint $\boldsymbol{\beta}_2 = 0$ is valid! This result is strongly counter-intuitive. It seems to contradict our initial motivation for pursuing constrained estimation – to improve estimation efficiency.

It turns out that a more refined answer is appropriate. Constrained estimation is desirable, but not constrained least-squares estimation. While least-squares is asymptotically efficient for estimation of the unconstrained projection model, it is not an efficient estimator of the constrained projection model.

8.10 Variance and Standard Error Estimation

We have discussed covariance matrix estimation for the CLS estimator, but not yet for the EMD estimator.

The asymptotic covariance matrix (8.29) may be estimated by replacing $\boldsymbol{V}_{\boldsymbol{\beta}}$ with a consistent estimate. It is best to construct the variance estimate using $\widetilde{\boldsymbol{\beta}}_{\text{emd}}$. The EMD residuals are $\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}_{\text{emd}}$. Using these we can estimate the matrix $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'e_i^2\right)$ by

$$\widetilde{\boldsymbol{\Omega}} = \frac{1}{n-k+q}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\widetilde{e}_i^2.$$

Following the formula for CLS we recommend an adjusted degrees of freedom. Given $\widetilde{\boldsymbol{\Omega}}$ the moment estimator of $\boldsymbol{V}_{\boldsymbol{\beta}}$ is

$$\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}} = \widehat{\boldsymbol{Q}}_{xx}^{-1}\widetilde{\boldsymbol{\Omega}}\widehat{\boldsymbol{Q}}_{xx}^{-1}$$

Given this, we construct the variance estimator

$$\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta},\text{emd}} = \widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}} - \widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}. \tag{8.38}$$

A standard error for $\boldsymbol{h}'\widetilde{\boldsymbol{\beta}}$ is then

$$s(\boldsymbol{h}'\widetilde{\boldsymbol{\beta}}) = \left(n^{-1}\boldsymbol{h}'\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta},\text{emd}}\boldsymbol{h}\right)^{1/2}. \tag{8.39}$$

8.11 Hausman Equality

Form (8.28) we have

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right) = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\sqrt{n}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{c}\right)$$

$$\xrightarrow{d} \text{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\right).$$

It follows that the asymptotic variances of the estimators satisfy the relationship

$$\operatorname{avar}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}-\widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)=\operatorname{avar}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}}\right)-\operatorname{avar}\left(\widetilde{\boldsymbol{\beta}}_{\text{emd}}\right). \tag{8.40}$$

We call (8.40) the Hausman Equality: the asymptotic variance of the difference between an efficient and inefficient estimator is the difference in the asymptotic variances.

8.12 Example: Mankiw, Romer and Weil (1992)

We illustrate the methods by replicating some of the estimates reported in a well-known paper by Mankiw, Romer, and Weil (1992). The paper investigates the implications of the Solow growth model using cross-country regressions. A key equation in their paper regresses the change between 1960 and 1985 in log GDP per capita on (1) log GDP in 1960, (2) the log of the ratio of aggregate investment to GDP, (3) the log of the sum of the population growth rate n, the technological growth rate g, and the rate of depreciation δ, and (4) the log of the percentage of the working-age population that is in secondary schoool (*School*), the latter a proxy for human-capital accumulation.

The data is available on the textbook webpage in the file `MRW1992`.

The sample is 98 non-oil-producing countries, and the data was reported in the published paper. As g and δ were unknown the authors set $g+\delta=0.05$. We report least-squares estimates in the first column of the table below, using the authors' original data. The estimates are consistent with the Solow theory due to the positive coefficients on investment and human capital and negative coefficient for population growth. The estimates are also consistent with the convergence hypothesis (that income levels tend towards a common mean over time) as the coefficient on intial GDP is negative.

The authors show that in the Solow model the 2^{nd}, 3^{rd} and 4^{th} coefficients sum to zero. They reestimated the equation imposing this contraint. We present constrained least-squares estimates in the second column, and efficient minimum distance estimates in the third column. Most of the coefficients and standard errors only exhibit small changes by imposing the constaint. The one exception is the coefficient on log population growth, which increases in magnitude and its standard error decreases substantially. The differences between the CLS and EMD estimates are modest but not inconsequential.

Table
Estimates of Solow Growth Model
Dependent Variable $\log\frac{GDP_{1985}}{GDP_{1960}}$

	$\widehat{\beta}_{OLS}$	$\widehat{\beta}_{CLS}$	$\widehat{\beta}_{EMD}$
$\log GDP_{1960}$	−0.29	−0.30	−0.30
	(0.05)	(0.05)	(0.05)
$\log\frac{I}{GDP}$	0.52	0.50	0.46
	(0.11)	(0.09)	(0.08)
$\log(n+g+\delta)$	−0.51	−0.74	−0.71
	(0.25)	(0.08)	(0.08)
$\log School$	0.23	0.24	0.25
	(0.07)	(0.07)	(0.07)
Intercept	3.02	2.46	2.48
	(0.74)	(0.44)	(0.44)

Note: Standard errors are heteroskedasticity-consistent

We now present Stata, R and MATLAB code which implements these estimates.

You may notice that the Stata code has a section which uses the Mata matrix programming language. This is used because Stata does not implement the efficient minimum distance estimator, so needs to be separately programmed. As illustrated here, the Mata language allows a Stata user to implement methods using commands which are quite similar to MATLAB.

Stata do File

```
use "MRW1992.dta", clear
gen lndY = log(Y85)-log(Y60)
gen lnY60 = log(Y60)
gen lnI = log(invest/100)
gen lnG = log(pop_growth/100+0.05)
gen lnS = log(school/100)
// Unrestricted regression
reg lndY lnY60 lnI lnG lnS if N==1, r
// Store result for efficient minimum distance
mat b = e(b)'
scalar k = e(rank)
mat V = e(V)
// Constrained regression
constraint define 1 lnI+lnG+lnS=0
cnsreg lndY lnY60 lnI lnG lnS if N==1, constraints(1) r
// Efficient minimum distance
mata{
      data = st_data(.,("lnY60","lnI","lnG","lnS","lndY","N"))
      data_select = select(data,data[.,6]:==1)
      y = data_select[.,5]
      n = rows(y)
      x = (data_select[.,1..4],J(n,1,1))
      k = cols(x)
      invx = invsym(x'*x)
      b_ols = st_matrix("b")
      V_ols = st_matrix("V")
      R = (0\1\1\1\0)
      b_emd = b_ols-V_ols*R*invsym(R'*V_ols*R)*R'*b_ols
      e_emd = J(1,k,y-x*b_emd)
      xe_emd = x:*e_emd
      xe_emd'*xe_emd
      V2 = (n/(n-k+1))*invx*(xe_emd'*xe_emd)*invx
      V_emd = V2 - V2*R*invsym(R'*V2*R)*R'*V2
      se_emd = diagonal(sqrt(V_emd))
      st_matrix("b_emd",b_emd)
      st_matrix("se_emd",se_emd)}
mat list b_emd
mat list se_emd
```

R Program File

```
# Load the data and create variables
data <- read.table("MRW1992.txt",header=TRUE)
N <- matrix(data$N,ncol=1)
lndY <- matrix(log(data$Y85)-log(data$Y60),ncol=1)
lnY60 <- matrix(log(data$Y60),ncol=1)
lnI <- matrix(log(data$invest/100),ncol=1)
lnG <- matrix(log(data$pop_growth/100+0.05),ncol=1)
lnS <- matrix(log(data$school/100),ncol=1)
xx <- as.matrix(cbind(lnY60,lnI,lnG,lnS,matrix(1,nrow(lndY),1)))
x <- xx[N==1,]
y <- lndY[N==1]
n <- nrow(x)
k <- ncol(x)
# Unrestricted regression
invx <-solve(t(x)%*%x)
beta_ols <- invx%*%t(x)%*%y
e_ols <- rep((y-x%*%beta_ols),times=k)
xe_ols <- x*e_ols
V_ols <- (n/(n-k))*invx%*%(t(xe_ols)%*%xe_ols)%*%invx
se_ols <- sqrt(diag(V_ols))
print(beta_ols)
print(se_ols)
# Constrained regression
R <- c(0,1,1,1,0)
iR = invx%*%R%*%solve(t(R)%*%invx%*%R)%*%t(R)
b_cls <- b_ols - iR%*%b_ols
e_cls <- rep((y-x%*%b_cls),times=k)
xe_cls <- x*e_cls
V_tilde <- (n/(n-k+1))*invx%*%(t(xe_cls)%*%xe_cls)%*%invx
V_cls <- V_tilde - iR%*%V_tilde - V_tilde%*%t(iR) +
iR%*%V_tilde%*%t(iR)
print(b_cls)
print(se_cls)
# Efficient minimum distance
Vr = V_ols%*%R%*%solve(t(R)%*%V_ols%*%R)%*%t(R)
b_emd <- b_ols - Vr%*%b_ols
e_emd <- rep((y-x%*%b_emd),times=k)
xe_emd <- x*e_emd
V2 <- (n/(n-k+1))*invx%*%(t(xe_emd)%*%xe_emd)%*%invx
V_emd <- V2 - V2%*%R%*%solve(t(R)%*%V2%*%R)%*%t(R)%*%V2
se_emd <- sqrt(diag(V_emd))
```

MATLAB Program File

```
% Load the data and create variables
data = xlsread('MRW1992.xlsx');
N = data(:,1);
Y60 = data(:,4);
Y85 = data(:,5);
pop_growth = data(:,7);
invest = data(:,8);
school = data(:,9);
lndY = log(Y85)-log(Y60);
lnY60 = log(Y60);
lnI = log(invest/100);
lnG = log(pop_growth/100+0.05);
lnS = log(school/100);
xx = [lnY60,lnI,lnG,lnS,ones(size(lndY,1),1)];
x = xx(N==1,:);
y = lndY(N==1);
[n,k] = size(x);
% Unrestricted regression
invx = inv(x'*x);
beta_ols = invx*x'*y;
e_ols = repmat((y-x*beta_ols),1,k);
xe_ols = x.*e_ols;
V_ols = (n/(n-k))*invx*(xe_ols'*xe_ols)*invx;
se_ols = sqrt(diag(V_ols));
display(beta_ols);
display(se_ols);
% Constrained regression
R = [0;1;1;1;0];
iR = invx*R*inv(R'*invx*R)*R';
beta_cls = beta_ols - iR*beta_ols;
e_cls = repmat((y-x*beta_cls),1,k);
xe_cls = x.*e_cls;
V_tilde = (n/(n-k+1))*invx*(xe_cls'*xe_cls)*invx;
V_cls = V_tilde - iR*V_tilde - V_tilde*(iR')...
+ iR*V_tilde*(iR');
se_cls = sqrt(diag(V_cls));
display(beta_cls);
display(se_cls);
% (3) Efficient minimum distance
beta_emd = beta_ols-V_ols*R*inv(R'*V_ols*R)*R'*beta_ols;
e_emd = repmat((y-x*beta_emd),1,k);
xe_emd = x.*e_emd;
V2 = (n/(n-k+1))*invx*(xe_emd'*xe_emd)*invx;
V_emd = V2 - V2*R*inv(R'*V2*R)*R'*V2;
se_emd = sqrt(diag(V_emd));
display(beta_emd);display(se_emd);
```

8.13 Misspecification

What are the consequences for a constrained estimator $\widetilde{\boldsymbol{\beta}}$ if the constraint (8.1) is incorrect? To be specific, suppose that

$$\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}^*$$

where $\boldsymbol{c}^*$ is not necessarily equal to $\boldsymbol{c}$.

This situation is a generalization of the analysis of "omitted variable bias" from Section 2.23, where we found that the short regression (e.g. (8.13)) is estimating a different projection coefficient than the long regression (e.g. (8.11)).

One mechanical answer is that we can use the formula (8.25) for the minimum distance estimator to find that

$$\widetilde{\boldsymbol{\beta}}_{\text{md}} \xrightarrow{p} \boldsymbol{\beta}^*_{\text{md}} = \boldsymbol{\beta} - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}(\boldsymbol{c}^* - \boldsymbol{c}). \tag{8.41}$$

The second term, $\boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}(\boldsymbol{c}^* - \boldsymbol{c})$, shows that imposing an incorrect constraint leads to inconsistency – an asymptotic bias. We can call the limiting value $\boldsymbol{\beta}^*_{\text{md}}$ the minimum-distance projection coefficient or the pseudo-true value implied by the restriction.

However, we can say more.

For example, we can describe some characteristics of the approximating projections. The CLS estimator projection coefficient has the representation

$$\boldsymbol{\beta}^*_{\text{cls}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}} \mathbb{E}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2,$$

the best linear predictor subject to the constraint (8.1). The minimum distance estimator converges to

$$\boldsymbol{\beta}^*_{\text{md}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \boldsymbol{W} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

where $\boldsymbol{\beta}_0$ is the true coefficient. That is, $\boldsymbol{\beta}^*_{\text{md}}$ is the coefficient vector satisfying (8.1) closest to the true value in the weighted Euclidean norm. These calculations show that the constrained estimators are still reasonable in the sense that they produce good approximations to the true coefficient, conditional on being required to satisfy the constraint.

We can also show that $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ has an asymptotic normal distribution. The trick is to define the pseudo-true value

$$\boldsymbol{\beta}^*_n = \boldsymbol{\beta} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}(\boldsymbol{c}^* - \boldsymbol{c}). \tag{8.42}$$

(Note that (8.41) and (8.42) are different!) Then

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\text{md}} - \boldsymbol{\beta}^*_n\right) &= \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\sqrt{n}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^*\right) \\
&= \left(\boldsymbol{I} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&\xrightarrow{d} \left(\boldsymbol{I} - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}) \\
&= \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})).
\end{aligned} \tag{8.43}$$

In particular

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\text{emd}} - \boldsymbol{\beta}^*_n\right) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}^*_{\boldsymbol{\beta}}).$$

This means that even when the constraint (8.1) is misspecified, the conventional covariance matrix estimator (8.38) and standard errors (8.39) are appropriate measures of the sampling variance, though the distributions are centered at the pseudo-true values (or projections) $\boldsymbol{\beta}^*_n$ rather than $\boldsymbol{\beta}$. The fact that the estimators are biased is an unavoidable consequence of misspecification.

An alternative approach to the asymptotic distribution theory under misspecification uses the concept of local alternatives. It is a technical device which might seem a bit artificial, but it is a powerful method to derive useful distributional approximations in a wide variety of contexts. The idea is to index the true coefficient $\boldsymbol{\beta}_n$ by n via the relationship

$$\boldsymbol{R}'\boldsymbol{\beta}_n = \boldsymbol{c} + \boldsymbol{\delta} n^{-1/2}. \tag{8.44}$$

Equation (8.44) specifies that $\boldsymbol{\beta}_n$ violates (8.1) and thus the constraint is misspecified. However, the constraint is "close" to correct, as the difference $\boldsymbol{R}'\boldsymbol{\beta}_n - \boldsymbol{c} = \boldsymbol{\delta} n^{-1/2}$ is "small" in the sense that it decreases with the sample size n. We call (8.44) **local misspecification**.

The asymptotic theory is then derived as $n \to \infty$ under the sequence of probability distributions with the coefficients $\boldsymbol{\beta}_n$. The way to think about this is that the true value of the parameter is $\boldsymbol{\beta}_n$, and it is "close" to satisfying (8.1). The reason why the deviation is proportional to $n^{-1/2}$ is because this is the only choice under which the localizing parameter $\boldsymbol{\delta}$ appears in the asymptotic distribution but does not dominate it. The best way to see this is to work through the asymptotic approximation.

Since $\boldsymbol{\beta}_n$ is the true coefficient value, then $y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_n + e_i$ and we have the standard representation for the unconstrained estimator, namely

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i e_i\right) \\ &\xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right).\end{aligned} \tag{8.45}$$

There is no difference under fixed (classical) or local asymptotics, since the right-hand-side is independent of the coefficient $\boldsymbol{\beta}_n$.

A difference arises for the constrained estimator. Using (8.44), $\boldsymbol{c} = \boldsymbol{R}'\boldsymbol{\beta}_n - \boldsymbol{\delta} n^{-1/2}$, so

$$\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c} = \boldsymbol{R}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) + \boldsymbol{\delta} n^{-1/2}$$

and

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} &= \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right) \\ &= \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) + \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta} n^{-1/2}.\end{aligned}$$

It follows that

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} - \boldsymbol{\beta}_n\right) = &\left(\boldsymbol{I} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_n\right) \\ &+ \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta}.\end{aligned}$$

The first term is asymptotically normal (from 8.45)). The second term converges in probability to a constant. This is because the $n^{-1/2}$ local scaling in (8.44) is exactly balanced by the $\sqrt{n}$ scaling of the estimator. No alternative rate would have produced this result.

Consequently, we find that the asymptotic distribution equals

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} - \boldsymbol{\beta}_n\right) &\xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right) + \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta} \\ &= \mathrm{N}\left(\boldsymbol{\delta}^*, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right)\end{aligned} \tag{8.46}$$

where

$$\boldsymbol{\delta}^* = \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta}.$$

The asymptotic distribution (8.46) is an approximation of the sampling distribution of the restricted estimator under misspecification. The distribution (8.46) contains an asymptotic bias component $\boldsymbol{\delta}^*$. The approximation is not fundamentally different from (8.43) – they both have the same asymptotic variances, and both reflect the bias due to misspecification. The difference is that (8.43) puts the bias on the left-side of the convergence arrow, while (8.46) has the bias on the right-side. There is no substantive difference between the two, but (8.46) is more convenient for some purposes, such as the analysis of the power of tests, as we will explore in the next chapter.

8.14 Nonlinear Constraints

In some cases it is desirable to impose nonlinear constraints on the parameter vector $\boldsymbol{\beta}$. They can be written as

$$\boldsymbol{r}(\boldsymbol{\beta}) = \mathbf{0} \tag{8.47}$$

where $\boldsymbol{r} : \mathbb{R}^k \to \mathbb{R}^q$. This includes the linear constraints (8.1) as a special case. An example of (8.47) which cannot be written as (8.1) is $\beta_1\beta_2 = 1$, which is (8.47) with $r(\boldsymbol{\beta}) = \beta_1\beta_2 - 1$.

The constrained least-squares and minimum distance estimators of $\boldsymbol{\beta}$ subject to (8.47) solve the minimization problems

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmin}} SSE(\boldsymbol{\beta}) \tag{8.48}$$

$$\widetilde{\boldsymbol{\beta}}_{\text{md}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmin}} \ J(\boldsymbol{\beta}) \tag{8.49}$$

where $SSE(\boldsymbol{\beta})$ and $J(\boldsymbol{\beta})$ are defined in (8.4) and (8.20), respectively. The solutions minimize the Lagrangians

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} SSE(\boldsymbol{\beta}) + \boldsymbol{\lambda}' \boldsymbol{r}(\boldsymbol{\beta}) \tag{8.50}$$

or

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} J(\boldsymbol{\beta}) + \boldsymbol{\lambda}' \boldsymbol{r}(\boldsymbol{\beta}) \tag{8.51}$$

over $(\boldsymbol{\beta}, \boldsymbol{\lambda})$.

Computationally, there is no general closed-form solution for the estimator so they must be found numerically. Algorithms to numerically solve (8.48) and (8.49) are known as **constrained optimization** methods, and are available in programming languages including MATLAB, GAUSS and R.

Assumption 8.14.1 $\boldsymbol{r}(\boldsymbol{\beta}) = \mathbf{0}$, $\boldsymbol{r}(\boldsymbol{\beta})$ *is continuously differentiable at the true* $\boldsymbol{\beta}$, *and* $\operatorname{rank}(\boldsymbol{R}) = q$, *where* $\boldsymbol{R} = \dfrac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\boldsymbol{\beta})'$.

The asymptotic distribution is a simple generalization of the case of a linear constraint, but the proof is more delicate.

Theorem 8.14.1 *Under Assumptions 7.1.2, 8.14.1, and 8.6.2, for* $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_{\text{md}}$ *and* $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_{\text{cls}}$ *defined in (8.48) and (8.49),*

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right)$$

as $n \to \infty$, *where* $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})$ *is defined in (8.27). For* $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$, $\boldsymbol{W} = \boldsymbol{Q}_{xx}$ *and* $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) = \boldsymbol{V}_{\text{cls}}$ *as defined in Theorem 8.6.3.* $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})$ *is minimized with* $\boldsymbol{W} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$, *in which case the asymptotic variance is*

$$\boldsymbol{V}_{\boldsymbol{\beta}}^{*} = \boldsymbol{V}_{\boldsymbol{\beta}} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}.$$

The asymptotic variance matrix for the efficient minimum distance estimator can be estimated by

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{*} = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} - \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$$

where

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}(\widetilde{\boldsymbol{\beta}}_{\text{md}})'. \tag{8.52}$$

Standard errors for the elements of $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ are the square roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widetilde{\boldsymbol{\beta}}}^{*} = n^{-1}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{*}$.

8.15 Inequality Restrictions

Inequality constraints on the parameter vector $\boldsymbol{\beta}$ take the form

$$\boldsymbol{r}(\boldsymbol{\beta}) \geq \mathbf{0} \tag{8.53}$$

for some function $\boldsymbol{r} : \mathbb{R}^k \to \mathbb{R}^q$. The most common example is a non-negative constraint

$$\beta_1 \geq 0.$$

The constrained least-squares and minimum distance estimators can be written as

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \underset{\boldsymbol{r}(\boldsymbol{\beta}) \geq \mathbf{0}}{\operatorname{argmin}}\, SSE(\boldsymbol{\beta}) \tag{8.54}$$

and

$$\widetilde{\boldsymbol{\beta}}_{\text{md}} = \underset{\boldsymbol{r}(\boldsymbol{\beta}) \geq \mathbf{0}}{\operatorname{argmin}}\; J\left(\boldsymbol{\beta}\right). \tag{8.55}$$

Except in special cases the constrained estimators do not have simple algebraic solutions. An important exception is when there is a single non-negativity constraint, e.g. $\beta_1 \geq 0$ with $q = 1$. In this case the constrained estimator can be found by two-step approach. First compute the uncontrained estimator $\widehat{\boldsymbol{\beta}}$. If $\widehat{\beta}_1 \geq 0$ then $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}$. Second, if $\widehat{\beta}_1 < 0$ then impose $\beta_1 = 0$ (eliminate the regressor X_1) and re-estimate. This yields the constrained least-squares estimator. While this method works when there is a single non-negativity constraint, it does not immediately generalize to other contexts.

The computational problems (8.54) and (8.55) are examples of **quadratic programming** problems. Quick and easy computer algorithms are available in programming languages including MATLAB, GAUSS and R.

Inference on inequality-constrained estimators is unfortunately quite challenging. The conventional asymptotic theory gives rise to the following dichotomy. If the true parameter satisfies the strict inequality $\boldsymbol{r}(\boldsymbol{\beta}) > \mathbf{0}$, then asymptotically the estimator is not subject to the constraint and the inequality-constrained estimator has an asymptotic distribution equal to the unconstrained case. However if the true parameter is on the boundary, e.g. $\boldsymbol{r}(\boldsymbol{\beta}) = \mathbf{0}$, then the estimator has a truncated structure. This is easiest to see in the one-dimensional case. If we have an estimator $\hat{\beta}$ which satisfies $\sqrt{n}\left(\widehat{\beta} - \beta\right) \xrightarrow{d} \mathrm{Z} = \mathrm{N}(0, V_\beta)$ and $\beta = 0$, then the constrained estimator $\widetilde{\beta} = \max[\widehat{\beta}, 0]$ will have the asymptotic distribution $\sqrt{n}\widetilde{\beta} \xrightarrow{d} \max[\mathrm{Z}, 0]$, a "half-normal" distribution.

8.16 Technical Proofs*

Proof of Theorem 8.8.1, Equation (8.31). Let $\boldsymbol{R}_\perp$ be a full rank $k \times (k-q)$ matrix satisfying $\boldsymbol{R}_\perp' \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{R} = \mathbf{0}$ and then set $\boldsymbol{C} = [\boldsymbol{R}, \boldsymbol{R}_\perp]$ which is full rank and invertible. Then we can calculate that

$$\begin{aligned}\boldsymbol{C}' \boldsymbol{V}_{\boldsymbol{\beta}}^* \boldsymbol{C} &= \begin{bmatrix} \boldsymbol{R}' \boldsymbol{V}_{\boldsymbol{\beta}}^* \boldsymbol{R} & \boldsymbol{R}' \boldsymbol{V}_{\boldsymbol{\beta}}^* \boldsymbol{R}_\perp \\ \boldsymbol{R}_\perp' \boldsymbol{V}_{\boldsymbol{\beta}}^* \boldsymbol{R} & \boldsymbol{R}_\perp' \boldsymbol{V}_{\boldsymbol{\beta}}^* \boldsymbol{R}_\perp \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}_\perp' \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{R}_\perp \end{bmatrix}\end{aligned}$$

and

$$\begin{aligned}&\boldsymbol{C}' \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) \boldsymbol{C} \\ &= \begin{bmatrix} \boldsymbol{R}' \boldsymbol{V}_{\boldsymbol{\beta}}^*(\boldsymbol{W}) \boldsymbol{R} & \boldsymbol{R}' \boldsymbol{V}_{\boldsymbol{\beta}}^*(\boldsymbol{W}) \boldsymbol{R}_\perp \\ \boldsymbol{R}_\perp' \boldsymbol{V}_{\boldsymbol{\beta}}^*(\boldsymbol{W}) \boldsymbol{R} & \boldsymbol{R}_\perp' \boldsymbol{V}_{\boldsymbol{\beta}}^*(\boldsymbol{W}) \boldsymbol{R}_\perp \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}_\perp' \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{R}_\perp + \boldsymbol{R}_\perp' \boldsymbol{W} \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}\right)^{-1} \boldsymbol{R}' \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}\right)^{-1} \boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}_\perp \end{bmatrix}.\end{aligned}$$

Thus

$$\begin{aligned}&\boldsymbol{C}' \left(\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) - \boldsymbol{V}_{\boldsymbol{\beta}}^*\right) \boldsymbol{C} \\ &= \boldsymbol{C}' \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) \boldsymbol{C} - \boldsymbol{C}' \boldsymbol{V}_{\boldsymbol{\beta}}^* \boldsymbol{C} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}_\perp' \boldsymbol{W} \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}\right)^{-1} \boldsymbol{R}' \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}\right)^{-1} \boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}_\perp \end{bmatrix} \\ &\geq 0\end{aligned}$$

Since $\boldsymbol{C}$ is invertible it follows that $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) - \boldsymbol{V}_{\boldsymbol{\beta}}^* \geq 0$ which is (8.31). ■

Proof of Theorem 8.14.1. We show the result for the minimum distance estimator $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_{\mathrm{md}}$, as the proof for the constrained least-squares estimator is similar. For simplicity we assume that the constrained estimator is consistent $\widetilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$. This can be shown with more effort, but requires a deeper treatment than appropriate for this textbook.

For each element $r_j(\boldsymbol{\beta})$ of the q-vector $\boldsymbol{r}(\boldsymbol{\beta})$, by the mean value theorem there exists a $\boldsymbol{\beta}_j^*$ on the line segment joining $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ such that

$$r_j(\widetilde{\boldsymbol{\beta}}) = r_j(\boldsymbol{\beta}) + \frac{\partial}{\partial \boldsymbol{\beta}} r_j(\boldsymbol{\beta}_j^*)' \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{8.56}$$

Let $\boldsymbol{R}_n^*$ be the $k \times q$ matrix

$$\boldsymbol{R}^* = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\beta}} r_1(\boldsymbol{\beta}_1^*) & \frac{\partial}{\partial \boldsymbol{\beta}} r_2(\boldsymbol{\beta}_2^*) & \cdots & \frac{\partial}{\partial \boldsymbol{\beta}} r_q(\boldsymbol{\beta}_q^*) \end{bmatrix}.$$

Since $\widetilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ it follows that $\boldsymbol{\beta}_j^* \xrightarrow{p} \boldsymbol{\beta}$, and by the CMT, $\boldsymbol{R}^* \xrightarrow{p} \boldsymbol{R}$. Stacking the (8.56), we obtain

$$\boldsymbol{r}(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{r}(\boldsymbol{\beta}) + \boldsymbol{R}^{*\prime}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).$$

Since $\boldsymbol{r}(\widetilde{\boldsymbol{\beta}}) = \mathbf{0}$ by construction and $\boldsymbol{r}(\boldsymbol{\beta}) = \mathbf{0}$ by Assumption 8.6.1, this implies

$$\mathbf{0} = \boldsymbol{R}^{*\prime}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{8.57}$$

The first-order condition for (8.51) is

$$\widehat{\boldsymbol{W}}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right) = \widehat{\boldsymbol{R}}\widetilde{\boldsymbol{\lambda}}.$$

where $\widehat{\boldsymbol{R}}$ is defined in (8.52).

Premultiplying by $\boldsymbol{R}^{*\prime}\widehat{\boldsymbol{W}}^{-1}$, inverting, and using (8.57), we find

$$\widetilde{\boldsymbol{\lambda}} = \left(\boldsymbol{R}^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}^{*\prime}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right) = \left(\boldsymbol{R}^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}^{*\prime}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).$$

Thus

$$\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{I} - \widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\left(\boldsymbol{R}_n^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}_n^{*\prime}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{8.58}$$

From Theorem 7.3.2 and Theorem 7.7.1 we find

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &= \left(\boldsymbol{I} - \widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\left(\boldsymbol{R}_n^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widetilde{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}_n^{*\prime}\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&\xrightarrow{d} \left(\boldsymbol{I} - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right) \\
&= \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right).
\end{aligned}$$

■

Exercises

Exercise 8.1 In the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, show directly from definition (8.3) that the CLS estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ subject to the constraint that $\boldsymbol{\beta}_2 = \mathbf{0}$ is the OLS regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$.

Exercise 8.2 In the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, show directly from definition (8.3) that the CLS estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, subject to the constraint that $\boldsymbol{\beta}_1 = \boldsymbol{c}$ (where $\boldsymbol{c}$ is some given vector) is the OLS regression of $\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{c}$ on $\boldsymbol{X}_2$.

Exercise 8.3 In the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, with $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ each $n \times k$, find the CLS estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$, subject to the constraint that $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2$.

Exercise 8.4 In the linear projection model $y_i = \alpha + \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$, consider the restriction $\boldsymbol{\beta} = \mathbf{0}$.

(a) Find the constrained least-squares (CLS) estimator of α under the restriction $\boldsymbol{\beta} = \mathbf{0}$.

(b) Find an expression for the efficient minimum distance estimator of α under the restriction $\boldsymbol{\beta} = \mathbf{0}$.

Exercise 8.5 Verify that for $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ defined in (8.9) that $\boldsymbol{R}'\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \boldsymbol{c}$.

Exercise 8.6 Prove Theorem 8.4.1

Exercise 8.7 Prove Theorem 8.4.2, that is, $\mathbb{E}\left(\widetilde{\boldsymbol{\beta}}_{\text{cls}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$, under the assumptions of the linear regression regression model and (8.1).

Hint: Use Theorem 8.4.1.

Exercise 8.8 Prove Theorem 8.4.3.

Exercise 8.9 Prove Theorem 8.4.4, that is, $\mathbb{E}\left(s^2_{\text{cls}} \mid \boldsymbol{X}\right) = \sigma^2$, under the assumptions of the homoskedastic regression model and (8.1).

Exercise 8.10 Verify (8.24) and (8.25), and that the minimum distance estimator $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ with $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{xx}$ equals the CLS estimator.

Exercise 8.11 Prove Theorem 8.6.1.

Exercise 8.12 Prove Theorem 8.6.2.

Exercise 8.13 Prove Theorem 8.6.3. (Hint: Use that CLS is a special case of Theorem 8.6.2.)

Exercise 8.14 Verify that (8.29) is $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})$ with $\boldsymbol{W} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$.

Exercise 8.15 Prove (8.30). Hint: Use (8.29).

Exercise 8.16 Verify (8.32), (8.33) and (8.34)

Exercise 8.17 Verify (8.35), (8.36), and (8.37).

Exercise 8.18 Suppose you have two independent samples

$$y_{1i} = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_{1i}$$

and

$$y_{2i} = \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_{2i}$$

both of sample size n, and both $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are $k \times 1$. You estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ by OLS on each sample, $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$, say, with asymptotic covariance matrix estimators $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}_1}$ and $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}_2}$ (which are consistent for the asymptotic covariance matrices $\boldsymbol{V}_{\boldsymbol{\beta}_1}$ and $\boldsymbol{V}_{\boldsymbol{\beta}_2}$). Consider efficient minimimum distance estimation under the restriction $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

(a) Find the estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} = \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$

(b) Find the asymptotic distribution of $\widetilde{\boldsymbol{\beta}}$.

(c) How would you approach the problem if the sample sizes are different, say n_1 and n_2?

Exercise 8.19 As in Exercise 7.29 and 3.24, use the CPS dataset and the subsample of white male Hispanics.

(a) Estimate the regression

$$\begin{aligned}\widehat{\log(Wage)} = {} & \beta_1\, education + \beta_2\, experience + \beta_3\, experience^2/100 + \beta_4 Married_1 \\ & + \beta_5 Married_2 + \beta_6 Married_3 + \beta_7 Widowed + \beta_8 Divorced + \beta_9 Separated + \beta_{10}\end{aligned}$$

where $Married_1$, $Married_2$, and $Married_3$ are the first three marital status codes as listed in Section 3.19.

(b) Estimate the equation using constrained least-squares, imposing the constraints $\beta_4 = \beta_7$ and $\beta_8 = \beta_9$, and report the estimates and standard errors

(c) Estimate the equation using efficient minimum distance, imposing the same constraints, and report the estimates and standard errors

(d) Under what constraint on the coefficients is the wage equation non-decreasing in experience for experience up to 50?

(e) Estimate the equation imposing $\beta_4 = \beta_7$, $\beta_8 = \beta_9$, and the inequality from part (d).

Exercise 8.20 Take the model

$$\begin{aligned} y_i &= m(x_i) + e_i \\ m(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p \\ \mathbb{E}\left(z_i e_i\right) &= 0 \\ z_i &= (1, x_i, ..., x_i^p)' \\ g(x) &= \frac{d}{dx} m(x) \end{aligned}$$

with iid observations (y_i, x_i), $i = 1, ..., n$. The order of the polynomial p is known.

(a) How should we interpret the function $m(x)$ given the projection assumption? How should we interpret $g(x)$? (Briefly)

(b) Describe an estimator $\widehat{g}(x)$ of $g(x)$.

(c) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{g}(x) - g(x)\right)$ as $n \to \infty$.

(d) Show how to construct an asymptotic 95% confidence interval for $g(x)$ (for a single x).

(e) Assume $p = 2$. Describe how to estimate $g(x)$ imposing the constraint that $m(x)$ is concave.

(f) Assume $p = 2$. Describe how to estimate $g(x)$ imposing the constraint that $m(u)$ is increasing on the region $u \in [x_L, x_U]$.

Exercise 8.21 Take the linear model with restrictions

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \\ \boldsymbol{R}'\boldsymbol{\beta} &= \boldsymbol{c} \end{aligned}$$

with n observations. Consider three estimators for $\boldsymbol{\beta}$

- $\widehat{\boldsymbol{\beta}}$, the unconstrained least squares estimator
- $\widetilde{\boldsymbol{\beta}}$, the constrained least squares estimator
- $\overline{\boldsymbol{\beta}}$, the constrained efficient minimum distance estimator

For each estimator, define the residuals $\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$, $\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}$, $\overline{e}_i = y_i - \boldsymbol{x}_i'\overline{\boldsymbol{\beta}}$, and variance estimators $\widehat{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^n \widehat{e}_i^2$, $\widetilde{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^n \widetilde{e}_i^2$, and $\overline{\sigma}^2 = \dfrac{1}{n}\sum_{i=1}^n \overline{e}_i^2$.

(a) As $\overline{\boldsymbol{\beta}}$ is the most efficient estimator and $\widehat{\boldsymbol{\beta}}$ the least, do you expect that $\overline{\sigma}^2 < \widetilde{\sigma}^2 < \widehat{\sigma}^2$, in large samples?

(b) Consider the statistic

$$T_n = \widehat{\sigma}^{-2}\sum_{i=1}^n \left(\widehat{e}_i - \widetilde{e}_i\right)^2$$

Find the asymptotic distribution for T_n when $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}$ is true.

(c) Does the result of the previous question simplify when the error e_i is homoskedastic?

Exercise 8.22 Take the linear model

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned}$$

with n observations. Consider the restriction

$$\frac{\beta_1}{\beta_2} = 2$$

(a) Find an explicit expression for the constrained least-squares (CLS) estimator $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \widetilde{\beta}_2)$ of $\beta = (\beta_1, \beta_2)$ under the restriction. Your answer should be specific to the restriction, it should not be a generic formula for an abstract general restriction.

(b) Derive the asymptotic distribution of $\widetilde{\beta}_1$ under the assumption that the restriction is true.

Chapter 9

Hypothesis Testing

In Chapter 5 we briefly introduced hypothesis testing in the context of the normal regression model. In this chapter we explore hypothesis testing in greater detail, with a particular emphasis on asymptotic inference.

9.1 Hypotheses

In Chapter 8 we discussed estimation subject to restrictions, including linear restrictions (8.1), nonlinear restrictions (8.47), and inequality restrictions (8.53). In this chapter we discuss **tests** of such restrictions.

Hypothesis tests attempt to assess whether there is evidence to contradict a proposed parametric restriction. Let

$$\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$$

be a $q \times 1$ parameter of interest where $\boldsymbol{r} : \mathbb{R}^k \to \Theta \subset \mathbb{R}^q$ is some transformation. For example, $\boldsymbol{\theta}$ may be a single coefficient, e.g. $\boldsymbol{\theta} = \beta_j$, the difference between two coefficients, e.g. $\boldsymbol{\theta} = \beta_j - \beta_\ell$, or the ratio of two coefficients, e.g. $\boldsymbol{\theta} = \beta_j / \beta_\ell$.

A point hypothesis concerning $\boldsymbol{\theta}$ is a proposed restriction such as

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 \tag{9.1}$$

where $\boldsymbol{\theta}_0$ is a hypothesized (known) value.

More generally, letting $\boldsymbol{\beta} \in \boldsymbol{B} \subset \mathbb{R}^k$ be the parameter space, a hypothesis is a restriction $\boldsymbol{\beta} \in \boldsymbol{B}_0$ where $\boldsymbol{B}_0$ is a proper subset of $\boldsymbol{B}$. This specializes to (9.1) by setting $\boldsymbol{B}_0 = \{\boldsymbol{\beta} \in \boldsymbol{B} : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0\}$.

In this chapter we will focus exclusively on point hypotheses of the form (9.1) as they are the most common and relatively simple to handle.

The hypothesis to be tested is called the null hypothesis.

Definition 9.1.1 *The* ***null hypothesis****, written* $\mathbb{H}_0$*, is the restriction* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ *or* $\boldsymbol{\beta} \in \boldsymbol{B}_0$.

We often write the null hypothesis as $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ or $\mathbb{H}_0 : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$.

The complement of the null hypothesis (the collection of parameter values which do not satisfy the null hypothesis) is called the alternative hypothesis.

Definition 9.1.2 *The* ***alternative hypothesis****, written* $\mathbb{H}_1$*, is the set* $\{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0\}$ *or* $\{\beta \in \boldsymbol{B} \colon \beta \notin \boldsymbol{B}_0\}$.

We often write the alternative hypothesis as $\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ or $\mathbb{H}_1 : \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$. For simplicity, we often refer to the hypotheses as "the null" and "the alternative".

In hypothesis testing, we assume that there is a true (but unknown) value of $\boldsymbol{\theta}$ and this value either satisfies $\mathbb{H}_0$ or does not satisfy $\mathbb{H}_0$. The goal of hypothesis testing is to assess whether or not $\mathbb{H}_0$ is true, by asking if $\mathbb{H}_0$ is consistent with the observed data.

To be specific, take our example of wage determination and consider the question: Does union membership affect wages? We can turn this into a hypothesis test by specifying the null as the restriction that a coefficient on union membership is zero in a wage regression. Consider, for example, the estimates reported in Table 4.1. The coefficient for "Male Union Member" is 0.095 (a wage premium of 9.5%) and the coefficient for "Female Union Member" is 0.022 (a wage premium of 2.2%). These are estimates, not the true values. The question is: Are the true coefficients zero? To answer this question, the testing method asks the question: Are the observed estimates compatible with the hypothesis, in the sense that the deviation from the hypothesis can be reasonably explained by stochastic variation? Or are the observed estimates incompatible with the hypothesis, in the sense that that the observed estimates would be highly unlikely if the hypothesis were true?

9.2 Acceptance and Rejection

A hypothesis test either accepts the null hypothesis or rejects the null hypothesis in favor of the alternative hypothesis. We can describe these two decisions as "Accept $\mathbb{H}_0$" and "Reject $\mathbb{H}_0$". In the example given in the previous section, the decision would be either to accept the hypothesis that union membership does not affect wages, or to reject the hypothesis in favor of the alternative that union membership does affect wages.

The decision is based on the data, and so is a mapping from the sample space to the decision set. This splits the sample space into two regions S_0 and S_1 such that if the observed sample falls into S_0 we accept $\mathbb{H}_0$, while if the sample falls into S_1 we reject $\mathbb{H}_0$. The set S_0 is called the **acceptance region** and the set S_1 the **rejection** or **critical region**.

It is convenient to express this mapping as a real-valued function called a **test statistic**

$$T = T\left((y_1, \boldsymbol{x}_1), ..., (y_n, \boldsymbol{x}_n)\right)$$

relative to a **critical value** c. The hypothesis test then consists of the decision rule

1. Accept $\mathbb{H}_0$ if $T \leq c$.
2. Reject $\mathbb{H}_0$ if $T > c$.

A test statistic T should be designed so that small values are likely when $\mathbb{H}_0$ is true and large values are likely when $\mathbb{H}_1$ is true. There is a well developed statistical theory concerning the design of optimal tests. We will not review that theory here, but instead refer the reader to Lehmann and Romano (2005). In this chapter we will summarize the main approaches to the design of test statistics.

The most commonly used test statistic is the absolute value of the t-statistic

$$T = |T(\theta_0)| \tag{9.2}$$

where

$$T(\theta) = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \tag{9.3}$$

is the t-statistic from (7.43), $\widehat{\theta}$ is a point estimate, and $s(\widehat{\theta})$ its standard error. T is an appropriate statistic when testing hypotheses on individual coefficients or real-valued parameters $\theta = h(\boldsymbol{\beta})$ and θ_0 is the hypothesized value. Quite typically, $\theta_0 = 0$, as interest focuses on whether or not a coefficient equals zero, but this is not the only possibility. For example, interest may focus on whether an elasticity θ equals 1, in which case we may wish to test $\mathbb{H}_0 : \theta = 1$.

9.3 Type I Error

A false rejection of the null hypothesis $\mathbb{H}_0$ (rejecting $\mathbb{H}_0$ when $\mathbb{H}_0$ is true) is called a **Type I error**. The probability of a Type I error is

$$\Pr\left(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ true}\right) = \Pr\left(T > c \mid \mathbb{H}_0 \text{ true}\right). \tag{9.4}$$

The finite sample **size** of the test is defined as the supremum of (9.4) across all data distributions which satisfy $\mathbb{H}_0$. A primary goal of test construction is to limit the incidence of Type I error by bounding the size of the test.

For the reasons discussed in Chapter 7, in typical econometric models the exact sampling distributions of estimators and test statistics are unknown and hence we cannot explicitly calculate (9.4). Instead, we typically rely on asymptotic approximations. Suppose that the test statistic has an asymptotic distribution under $\mathbb{H}_0$. That is, when $\mathbb{H}_0$ is true

$$T \xrightarrow{d} \xi \tag{9.5}$$

as $n \to \infty$ for some continuously-distributed random variable ξ. This is not a substantive restriction, as most conventional econometric tests satisfy (9.5). Let $G(u) = \Pr(\xi \leq u)$ denote the distribution of ξ. We call ξ (or G) the **asymptotic null distribution**.

It is generally desirable to design test statistics T whose asymptotic null distribution G is known and does not depend on unknown parameters. In this case we say that the statistic T is **asymptotically pivotal**.

For example, if the test statistic equals the absolute t-statistic from (9.2), then we know from Theorem 7.12.1 that if $\theta = \theta_0$ (that is, the null hypothesis holds), then $T \xrightarrow{d} |\mathrm{Z}|$ as $n \to \infty$ where $\mathrm{Z} \sim \mathrm{N}(0,1)$. This means that $G(u) = \Pr(|\mathrm{Z}| \leq u) = 2\Phi(u) - 1$, the distribution of the absolute value of the standard normal as shown in (7.44). This distribution does not depend on unknowns and is pivotal.

We define the **asymptotic size** of the test as the asymptotic probability of a Type I error:

$$\begin{aligned}\lim_{n\to\infty} \Pr\left(T > c \mid \mathbb{H}_0 \text{ true}\right) &= \Pr\left(\xi > c\right) \\ &= 1 - G(c).\end{aligned}$$

We see that the asymptotic size of the test is a simple function of the asymptotic null distribution G and the critical value c. For example, the asymptotic size of a test based on the absolute t-statistic with critical value c is $2\left(1 - \Phi(c)\right)$.

In the dominant approach to hypothesis testing, the researcher pre-selects a **significance level** $\alpha \in (0,1)$ and then selects c so that the (asymptotic) size is no larger than α. When the asymptotic null distribution G is pivotal, we can accomplish this by setting c equal to the $1-\alpha$ quantile of the distribution G. (If the distribution G is not pivotal, more complicated methods must be used, pointing out the great convenience of using asymptotically pivotal test statistics.) We call c the **asymptotic critical value** because it has been selected from the asymptotic null distribution. For example, since $2\left(1 - \Phi(1.96)\right) = 0.05$, it follows that the 5% asymptotic critical value for the absolute t-statistic is $c = 1.96$. Calculation of normal critical values is done numerically in statistical software. For example, in MATLAB the command is `norminv(1-`α`/2)`.

9.4 t tests

As we mentioned earlier, the most common test of the one-dimensional hypothesis

$$\mathbb{H}_0 : \theta = \theta_0 \tag{9.6}$$

against the alternative

$$\mathbb{H}_1 : \theta \neq \theta_0 \tag{9.7}$$

is the absolute value of the t-statistic (9.3). We now formally state its asymptotic null distribution, which is a simple application of Theorem 7.12.1.

Theorem 9.4.1 *Under Assumptions 7.1.2, 7.10.1, and* $\mathbb{H}_0 : \theta = \theta_0$,

$$T(\theta_0) \xrightarrow{d} \mathrm{Z}.$$

For c *satisfying* $\alpha = 2\left(1 - \Phi(c)\right)$,

$$\Pr\left(|T(\theta_0)| > c \mid \mathbb{H}_0\right) \longrightarrow \alpha,$$

and the test "Reject $\mathbb{H}_0$ *if* $|T(\theta_0)| > c$*" has asymptotic size* α.

The theorem shows that asymptotic critical values can be taken from the normal distribution. As in our discussion of asymptotic confidence intervals (Section 7.13), the critical value could alternatively be taken from the student t distribution, which would be the exact test in the normal regression model (Section 5.14). Indeed, t critical values are the default in packages such as Stata. Since the critical values from the student t distribution are (slightly) larger than those from the normal distribution, using student t critical values decreases the rejection probability of the test. In practical applications the difference is typically unimportant unless the sample size is quite small (in which case the asymptotic approximation should be questioned as well).

The alternative hypothesis $\theta \neq \theta_0$ is sometimes called a "two-sided" alternative. In contrast, sometimes we are interested in testing for one-sided alternatives such as

$$\mathbb{H}_1 : \theta > \theta_0 \tag{9.8}$$

or

$$\mathbb{H}_1 : \theta < \theta_0. \tag{9.9}$$

Tests of $\theta = \theta_0$ against $\theta > \theta_0$ or $\theta < \theta_0$ are based on the signed t-statistic $T = T(\theta_0)$. The hypothesis $\theta = \theta_0$ is rejected in favor of $\theta > \theta_0$ if $T > c$ where c satisfies $\alpha = 1 - \Phi(c)$. Negative values of T are not taken as evidence against $\mathbb{H}_0$, as point estimates $\widehat{\theta}$ less than θ_0 do not point to $\theta > \theta_0$. Since the critical values are taken from the single tail of the normal distribution, they are smaller than for two-sided tests. Specifically, the asymptotic 5% critical value is $c = 1.645$. Thus, we reject $\theta = \theta_0$ in favor of $\theta > \theta_0$ if $T > 1.645$.

Conversely, tests of $\theta = \theta_0$ against $\theta < \theta_0$ reject $\mathbb{H}_0$ for negative t-statistics, e.g. if $T \leq -c$. For this alternative large positive values of T are not evidence against $\mathbb{H}_0$. An asymptotic 5% test rejects if $T < -1.645$.

There seems to be an ambiguity. Should we use the two-sided critical value 1.96 or the one-sided critical value 1.645? The answer is that we should use one-sided tests and critical values only when the parameter space is known to satisfy a one-sided restriction such as $\theta \geq \theta_0$. This is when the test of $\theta = \theta_0$ against $\theta > \theta_0$ makes sense. If the restriction $\theta \geq \theta_0$ is not known *a priori*, then imposing this restriction to test $\theta = \theta_0$ against $\theta > \theta_0$ does not makes sense. Since linear regression coefficients typically do not have *a priori* sign restrictions, the standard convention is to use two-sided critical values.

This may seem contrary to the way testing is presented in statistical textbooks, which often focus on one-sided alternative hypotheses. The latter focus is primarily for pedagogy, as the one-sided theoretical problem is cleaner and easier to understand.

9.5 Type II Error and Power

A false acceptance of the null hypothesis $\mathbb{H}_0$ (accepting $\mathbb{H}_0$ when $\mathbb{H}_1$ is true) is called a **Type II error**. The rejection probability under the alternative hypothesis is called the **power** of the test, and equals 1 minus the probability of a Type II error:

$$\pi(\boldsymbol{\theta}) = \Pr\left(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_1 \text{ true}\right) = \Pr\left(T > c \mid \mathbb{H}_1 \text{ true}\right).$$

We call $\pi(\boldsymbol{\theta})$ the **power function** and is written as a function of $\boldsymbol{\theta}$ to indicate its dependence on the true value of the parameter $\boldsymbol{\theta}$.

In the dominant approach to hypothesis testing, the goal of test construction is to have high power subject to the constraint that the size of the test is lower than the pre-specified significance level. Generally, the power of a test depends on the true value of the parameter $\boldsymbol{\theta}$, and for a well behaved test the power is increasing both as $\boldsymbol{\theta}$ moves away from the null hypothesis $\boldsymbol{\theta}_0$ and as the sample size n increases.

Given the two possible states of the world ($\mathbb{H}_0$ or $\mathbb{H}_1$) and the two possible decisions (Accept $\mathbb{H}_0$ or Reject $\mathbb{H}_0$), there are four possible pairings of states and decisions as is depicted in the following chart.

Hypothesis Testing Decisions

	Accept $\mathbb{H}_0$	Reject $\mathbb{H}_0$
$\mathbb{H}_0$ true	Correct Decision	Type I Error
$\mathbb{H}_1$ true	Type II Error	Correct Decision

Given a test statistic T, increasing the critical value c increases the acceptance region S_0 while decreasing the rejection region S_1. This decreases the likelihood of a Type I error (decreases the size) but increases the likelihood of a Type II error (decreases the power). Thus the choice of c involves a trade-off between size and the power. This is why the significance level α of the test cannot be set arbitrarily small. (Otherwise the test will not have meaningful power.)

It is important to consider the power of a test when interpreting hypothesis tests, as an overly narrow focus on size can lead to poor decisions. For example, it is easy to design a test which has perfect size yet has trivial power. Specifically, for any hypothesis we can use the following test: Generate a random variable $U \sim U[0,1]$ and reject $\mathbb{H}_0$ if $U < \alpha$. This test has exact size of α. Yet the test also has power precisely equal to α. When the power of a test equals the size, we say that the test has **trivial power**. Nothing is learned from such a test.

9.6 Statistical Significance

Testing requires a pre-selected choice of significance level α, yet there is no objective scientific basis for choice of α. Nevertheless the common practice is to set $\alpha = 0.05$ (5%). Alternative values are $\alpha = 0.10$ (10%) and $\alpha = 0.01$ (1%). These choices are somewhat the by-product of traditional tables of critical values and statistical software.

The informal reasoning behind the choice of a 5% critical value is to ensure that Type I errors should be relatively unlikely – that the decision "Reject $\mathbb{H}_0$" has scientific strength – yet the test retains power against reasonable alternatives. The decision "Reject $\mathbb{H}_0$" means that the evidence is inconsistent with the null hypothesis, in the sense that it is relatively unlikely (1 in 20) that data generated by the null hypothesis would yield the observed test result.

In contrast, the decision "Accept $\mathbb{H}_0$" is not a strong statement. It does not mean that the evidence supports $\mathbb{H}_0$, only that there is insufficient evidence to reject $\mathbb{H}_0$. Because of this, it is more accurate to use the label "Do not Reject $\mathbb{H}_0$" instead of "Accept $\mathbb{H}_0$".

When a test rejects $\mathbb{H}_0$ at the 5% significance level it is common to say that the statistic is **statistically significant** and if the test accepts $\mathbb{H}_0$ it is common to say that the statistic is **not statistically significant** or that it is **statistically insignificant**. It is helpful to remember that this is simply a compact way of saying "Using the statistic T, the hypothesis $\mathbb{H}_0$ can [cannot] be rejected at the asymptotic 5% level." Furthermore, when the null hypothesis $\mathbb{H}_0 : \theta = 0$ is rejected it is common to say that the coefficient θ is statistically significant, because the test has rejected the hypothesis that the coefficient is equal to zero.

Let us return to the example about the union wage premium as measured in Table 4.1. The absolute t-statistic for the coefficient on "Male Union Member" is $0.095/0.020 = 4.7$, which is greater than the 5% asymptotic critical value of 1.96. Therefore we reject the hypothesis that union membership does not affect wages for men. In this case, we can say that union membership is statistically significant for men. However, the absolute t-statistic for the coefficient on "Female Union Member" is $0.023/0.020 = 1.2$, which is less than 1.96 and therefore we do not reject the hypothesis that union membership does not affect wages for women. In this case we find that membership for women is not statistically significant.

When a test accepts a null hypothesis (when a test is not statistically significant) a common misinterpretation is that this is evidence that the null hypothesis is true. This is incorrect. Failure to reject is by itself not evidence. Without an analysis of power, we do not know the likelihood of making a Type II error, and thus are uncertain. In our wage example, it would be a mistake to write that "the regression finds that female union membership has no effect on wages". This is an incorrect and most unfortunate interpretation. The test has failed to reject the hypothesis that the coefficient is zero, but that does not mean that the coefficient is actually zero.

When a test rejects a null hypothesis (when a test is statistically significant) it is strong evidence against the hypothesis (since if the hypothesis were true then rejection is an unlikely event). Rejection should be taken as evidence against the null hypothesis. However, we can never conclude that the null hypothesis is indeed false, as we cannot exclude the possibility that we are making a Type I error.

Perhaps more importantly, there is an important distinction between statistical and economic significance. If we correctly reject the hypothesis $\mathbb{H}_0 : \theta = 0$ it means that the true value of θ is non-zero. This includes the possibility that θ may be non-zero but close to zero in magnitude. This only makes sense if we interpret the parameters in the context of their relevant models. In our wage regression example, we might consider wage effects of 1% magnitude or less as being "close to zero". In a log wage regression this corresponds to a dummy variable with a coefficient less than 0.01. If the standard error is sufficiently small (less than 0.005) then a coefficient estimate of 0.01 will be statistically significant, but not economically significant. This occurs frequently in applications with very large sample sizes where standard errors can be quite small.

The solution is to focus whenever possible on confidence intervals and the economic meaning of the coefficients. For example, if the coefficient estimate is 0.005 with a standard error of 0.002 then a 95% confidence interval would be [0.001, 0.009] indicating that the true effect is likely between 0% and 1%, and hence is slightly positive but small. This is much more informative than the misleading statement "the effect is statistically positive".

9.7 P-Values

Continuing with the wage regression estimates reported in Table 4.1, consider another question: Does marriage status affect wages? To test the hypothesis that marriage status has no effect on wages, we examine the t-statistics for the coefficients on "Married Male" and "Married Female" in Table 4.1, which are $0.211/0.010 = 22$ and $0.016/0.010 = 1.7$, respectively. The first exceeds the asymptotic 5% critical value of 1.96, so we reject the hypothesis for men, though not for women. But the statistic for men is exceptionally high, and that for women is only slightly below the critical value. Suppose in contrast that the t-statistic had been 2.0, which is more than the critical

value. This would lead to the decision "Reject $\mathbb{H}_0$" rather than "Accept $\mathbb{H}_0$". Should we really be making a different decision if the t-statistic is 1.7 rather than 2.0? The difference in values is small, shouldn't the difference in the decision be also small? Thinking through these examples it seems unsatisfactory to simply report "Accept $\mathbb{H}_0$" or "Reject $\mathbb{H}_0$". These two decisions do not summarize the evidence. Instead, the magnitude of the statistic T suggests a "degree of evidence" against $\mathbb{H}_0$. How can we take this into account?

The answer is to report what is known as the **asymptotic p-value**

$$p = 1 - G(T).$$

Since the distribution function G is monotonically increasing, the p-value is a monotonically decreasing function of T and is an equivalent test statistic. Instead of rejecting $\mathbb{H}_0$ at the significance level α if $T > c$, we can reject $\mathbb{H}_0$ if $p < \alpha$. Thus it is sufficient to report p, and let the reader decide. In practice, the p-value is calculated numerically. For example, in MATLAB the command is `2*(1-normalcdf(abs(t)))`.

In is instructive to interpret p as the **marginal significance level**: the largest value of α for which the test T "rejects" the null hypothesis. That is, $p = 0.11$ means that T rejects $\mathbb{H}_0$ for all significance levels greater than 0.11, but fails to reject $\mathbb{H}_0$ for significance levels less than 0.11.

Furthermore, the asymptotic p-value has a very convenient asymptotic null distribution. Since $T \xrightarrow{d} \xi$ under $\mathbb{H}_0$, then $p = 1 - G(T) \xrightarrow{d} 1 - G(\xi)$, which has the distribution

$$\begin{aligned}
\Pr\left(1 - G(\xi) \leq u\right) &= \Pr\left(1 - u \leq G(\xi)\right) \\
&= 1 - \Pr\left(\xi \leq G^{-1}(1-u)\right) \\
&= 1 - G\left(G^{-1}(1-u)\right) \\
&= 1 - (1-u) \\
&= u,
\end{aligned}$$

which is the uniform distribution on $[0, 1]$. (This calculation assumes that $G(u)$ is strictly increasing which is true for conventional asymptotic distributions such as the normal.) Thus $p \xrightarrow{d} \mathrm{U}[0, 1]$. This means that the "unusualness" of p is easier to interpret than the "unusualness" of T.

An important caveat is that the p-value p should not be interpreted as the probability that either hypothesis is true. A common mis-interpretation is that p is the probability "that the null hypothesis is true." This is incorrect. Rather, p is the marginal significance level – a measure of the strength of information against the null hypothesis.

For a t-statistic, the p-value can be calculated either using the normal distribution or the student t distribution, the latter presented in Section 5.14. p-values calculated using the student t will be slightly larger, though the difference is small when the sample size is large.

Returning to our empirical example, for the test that the coefficient on "Married Male" is zero, the p-value is 0.000. This means that it would be nearly impossible to observe a t-statistic as large as 22 when the true value of the coefficient is zero. When presented with such evidence we can say that we "strongly reject" the null hypothesis, that the test is "highly significant", or that "the test rejects at any conventional critical value". In contrast, the p-value for the coefficient on "Married Female" is 0.094. In this context it is typical to say that the test is "close to significant", meaning that the p-value is larger than 0.05, but not too much larger.

A related (but somewhat inferior) empirical practice is to append asterisks (*) to coefficient estimates or test statistics to indicate the level of significance. A common practice to to append a single asterisk (*) for an estimate or test statistic which exceeds the 10% critical value (i.e., is significant at the 10% level), append a double asterisk (**) for a test which exceeds the 5% critical value, or append a triple asterisk (***) for a test which exceeds the 1% critical value. Such a practice can be better than a table of raw test statistics as the asterisks permit a quick interpretation of significance. On the other hand, asterisks are inferior to p-values, which are also easy and quick to

interpret. The goal is essentially the same; it seems wiser to report p-values whenever possible and avoid the use of asterisks.

Our recommendation is that the best empirical practice is to compute and report the asymptotic p-value p rather than simply the test statistic T, the binary decision Accept/Reject, or appending asterisks. The p-value is a simple statistic, easy to interpret, and contains more information than the other choices.

We now summarize the main features of hypothesis testing.

1. Select a significance level α.
2. Select a test statistic T with asymptotic distribution $T \xrightarrow{d} \xi$ under $\mathbb{H}_0$.
3. Set the asymptotic critical value c so that $1 - G(c) = \alpha$, where G is the distribution function of ξ.
4. Calculate the asymptotic p-value $p = 1 - G(T)$.
5. Reject $\mathbb{H}_0$ if $T > c$, or equivalently $p < \alpha$.
6. Accept $\mathbb{H}_0$ if $T \leq c$, or equivalently $p \geq \alpha$.
7. Report p to summarize the evidence concerning $\mathbb{H}_0$ versus $\mathbb{H}_1$.

9.8 t-ratios and the Abuse of Testing

In Section 4.18, we argued that a good applied practice is to report coefficient estimates $\widehat{\theta}$ and standard errors $s(\widehat{\theta})$ for all coefficients of interest in estimated models. With $\widehat{\theta}$ and $s(\widehat{\theta})$ the reader can easily construct confidence intervals $[\widehat{\theta} \pm 2s(\widehat{\theta})]$ and t-statistics $\left(\widehat{\theta} - \theta_0\right)/s(\widehat{\theta})$ for hypotheses of interest.

Some applied papers (especially older ones) report t-ratios $T = \widehat{\theta}/s(\widehat{\theta})$ instead of standard errors. This is poor econometric practice. While the same information is being reported (you can back out standard errors by division, e.g. $s(\widehat{\theta}) = \widehat{\theta}/T$), standard errors are generally more helpful to readers than t-ratios. Standard errors help the reader focus on the estimation precision and confidence intervals, while t-ratios focus attention on statistical significance. While statistical significance is important, it is less important that the parameter estimates themselves and their confidence intervals. The focus should be on the meaning of the parameter estimates, their magnitudes, and their interpretation, not on listing which variables have significant (e.g. non-zero) coefficients. In many modern applications, sample sizes are very large so standard errors can be very small. Consequently t-ratios can be large even if the coefficient estimates are economically small. In such contexts it may not be interesting to announce "The coefficient is non-zero!" Instead, what is interesting to announce is that "The coefficient estimate is economically interesting!"

In particular, some applied papers report coefficient estimates and t-ratios, and limit their discussion of the results to describing which variables are "significant" (meaning that their t-ratios exceed 2) and the signs of the coefficient estimates. This is very poor empirical work, and should be studiously avoided. It is also a recipe for banishment of your work to lower tier economics journals.

Fundamentally, the common t-ratio is a test for the hypothesis that a coefficient equals zero. This should be reported and discussed when this is an interesting economic hypothesis of interest. But if this is not the case, it is distracting.

One problem is that standard packages, such as Stata, by default report t-statistics and p-values for every estimated coefficient. While this can be useful (as a user doesn't need to explicitly ask to test an desired coefficient) it can be misleading as it may unintentionally suggest that the entire list of t-statistics and p-values are important. Instead, a user should focus on tests of scientifically motivated hypotheses.

In general, when a coefficient θ is of interest, it is constructive to focus on the point estimate, its standard error, and its confidence interval. The point estimate gives our "best guess" for the value. The standard error is a measure of precision. The confidence interval gives us the range of values consistent with the data. If the standard error is large then the point estimate is not a good summary about θ. The endpoints of the confidence interval describe the bounds on the likely possibilities. If the confidence interval embraces too broad a set of values for θ, then the dataset is not sufficiently informative to render useful inferences about θ. On the other hand if the confidence interval is tight, then the data have produced an accurate estimate, and the focus should be on the value and interpretation of this estimate. In contrast, the statement "the t-ratio is highly significant" has little interpretive value.

The above discussion requires that the researcher knows what the coefficient θ means (in terms of the economic problem) and can interpret values and magnitudes, not just signs. This is critical for good applied econometric practice.

For example, consider the question about the effect of marriage status on mean log wages. We had found that the effect is "highly significant" for men and "close to significant" for women. Now, let's construct asymptotic 95% confidence intervals for the coefficients. The one for men is [0.19, 0.23] and that for women is [−0.00, 0.03]. This shows that average wages for married men are about 19-23% higher than for unmarried men, which is substantial, while the difference for women is about 0-3%, which is small. These *magnitudes* are more informative than the results of the hypothesis tests.

9.9 Wald Tests

The t-test is appropriate when the null hypothesis is a real-valued restriction. More generally, there may be multiple restrictions on the coefficient vector $\boldsymbol{\beta}$. Suppose that we have $q > 1$ restrictions which can be written in the form (9.1). It is natural to estimate $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$ by the plug-in estimate $\widehat{\boldsymbol{\theta}} = \boldsymbol{r}(\widehat{\boldsymbol{\beta}})$. To test $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ one approach is to measure the magnitude of the discrepancy $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$. As this is a vector, there is more than one measure of its length. One simple measure is the weighted quadratic form known as the **Wald statistic**. This is (7.47) evaluated at the null hypothesis

$$W = W(\boldsymbol{\theta}_0) = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \tag{9.10}$$

where $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}$ is an estimate of $\boldsymbol{V}_{\widehat{\boldsymbol{\theta}}}$ and $\widehat{\boldsymbol{R}} = \dfrac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\widehat{\boldsymbol{\beta}})'$. Notice that we can write W alternatively as

$$W = n \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$$

using the asymptotic variance estimate $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$, or we can write it directly as a function of $\widehat{\boldsymbol{\beta}}$ as

$$W = \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}\right)^{-1} \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right). \tag{9.11}$$

Also, when $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta}$ is a linear function of $\boldsymbol{\beta}$, then the Wald statistic simplifies to

$$W = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right).$$

The Wald statistic W is a weighted Euclidean measure of the length of the vector $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$. When $q = 1$ then $W = T^2$, the square of the t-statistic, so hypothesis tests based on W and $|T|$ are equivalent. The Wald statistic (9.10) is a generalization of the t-statistic to the case of multiple restrictions. As the Wald statistic is symmetric in the argument $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ it treats positive and negative alternatives symmetrically. Thus the inherent alternative is always two-sided.

As shown in Theorem 7.16.1, when $\boldsymbol{\beta}$ satisfies $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ then $W \xrightarrow{d} \chi^2_q$, a chi-square random variable with q degrees of freedom. Let $G_q(u)$ denote the χ^2_q distribution function. For a given significance level α, the asymptotic critical value c satisfies $\alpha = 1 - G_q(c)$. For example, the 5% critical values for $q = 1$, $q = 2$, and $q = 3$ are 3.84, 5.99, and 7.82, respectively, and in general the level α critical value can be calculated in MATLAB as `chi2inv(1-α,q)`. An asymptotic test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $W > c$. As with t-tests, it is conventional to describe a Wald test as "significant" if W exceeds the 5% asymptotic critical value.

Theorem 9.9.1 *Under Assumptions 7.1.2 and 7.10.1, and* $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, *then*

$$W \xrightarrow{d} \chi^2_q,$$

and for c *satisfying* $\alpha = 1 - G_q(c)$,

$$\Pr(W > c \mid \mathbb{H}_0) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ *if* $W > c$*" has asymptotic size* α.

Notice that the asymptotic distribution in Theorem 9.9.1 depends solely on q, the number of restrictions being tested. It does not depend on k, the number of parameters estimated.

The asymptotic p-value for W is $p = 1 - G_q(W)$, and this is particularly useful when testing multiple restrictions. For example, if you write that a Wald test on eight restrictions ($q = 8$) has the value $W = 11.2$, it is difficult for a reader to assess the magnitude of this statistic unless they have quick access to a statistical table or software. Instead, if you write that the p-value is $p = 0.19$ (as is the case for $W = 11.2$ and $q = 8$) then it is simple for a reader to interpret its magnitude as "insignificant". To calculate the asymptotic p-value for a Wald statistic in MATLAB, use the command `1-chi2cdf(w,q)`.

Some packages (including Stata) and papers report F versions of Wald statistics. That is, for any Wald statistic W which tests a q-dimensional restriction, the F version of the test is

$$F = W/q.$$

When F is reported, it is conventional to use $F_{q,n-k}$ critical values and p-values rather than χ^2_q values. The connection between Wald and F statistics is demonstrated in Section 9.14 we show that when Wald statistics are calculated using a homoskedastic covariance matrix, then $F = W/q$ is identicial to the F statistic of (5.23). While there is no formal justification to using the $F_{q,n-k}$ distribution for non-homoskedastic covariance matrices, the $F_{q,n-k}$ distribution provides continuity with the exact distribution theory under normality and is a bit more conservative than the χ^2_q distribution. (Furthermore, the difference is small when $n - k$ is moderately large.)

To implement a test of zero restrictions in Stata, an easy method is to use the command "test X1 X2" where X1 and X2 are the names of the variables whose coefficients are hypothesized to equal zero. This command should be executed after executing a regression command. The F version of the Wald statistic is reported, using the covariance matrix calculated using the method specified in the regression command. A p-value is reported, calculated using the $F_{q,n-k}$ distribution.

To illustrate, consider the empirical results presented in Table 4.1. The hypothesis "Union membership does not affect wages" is the joint restriction that both coefficients on "Male Union Member" and "Female Union Member" are zero. We calculate the Wald statistic for this joint hypothesis and find $W = 23$ (or $F = 12.5$) with a p-value of $p = 0.000$. Thus we reject the null hypothesis in favor of the alternative that at least one of the coefficients is non-zero. This does not mean that both coefficients are non-zero, just that one of the two is non-zero. Therefore examining both the joint Wald statistic and the individual t-statistics is useful for interpretation.

As a second example from the same regression, take the hypothesis that married status has no effect on mean wages for women. This is the joint restriction that the coefficients on "Married Female" and "Formerly Married Female" are zero. The Wald statistic for this hypothesis is $W = 6.4$ ($F = 3.2$) with a p-value of 0.04. Such a p-value is typically called "marginally significant", in the sense that it is slightly smaller than 0.05.

Abraham Wald

The Hungarian mathematician/statistician/econometrician Abraham Wald (1902-1950) developed an optimality property for the Wald test in terms of weighted average power. He also developed the field of sequential testing and the design of experiments.

9.10 Homoskedastic Wald Tests

If the error is known to be homoskedastic, then it is appropriate to use the homoskedastic Wald statistic (7.49) which replaces $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}$ with the homoskedastic estimate $\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\theta}}}$. This statistic equals

$$\begin{aligned} W^0 &= \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\theta}}}\right)^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\ &= \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \widehat{\boldsymbol{R}}\right)^{-1} \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right) / s^2. \end{aligned} \tag{9.12}$$

In the case of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$ we can write this as

$$W^0 = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right) / s^2. \tag{9.13}$$

We call (9.12) or (9.13) a **homoskedastic Wald statistic** as it is an appropriate test when the errors are conditionally homoskedastic.

As for W, when $q = 1$ then $W^0 = T^2$, the square of the t-statistic where the latter is computed with a homoskedastic standard error.

Theorem 9.10.1 *Under Assumptions 7.1.2 and 7.10.1,* $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, *and* $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, *then*

$$W^0 \xrightarrow{d} \chi^2_q,$$

and for c *satisfying* $\alpha = 1 - G_q(c)$,

$$\Pr\left(W^0 > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ *if* $W^0 > c$*" has asymptotic size* α.

9.11 Criterion-Based Tests

The Wald statistic is based on the length of the vector $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$: the discrepancy between the estimate $\widehat{\boldsymbol{\theta}} = \boldsymbol{r}(\widehat{\boldsymbol{\beta}})$ and the hypothesized value $\boldsymbol{\theta}_0$. An alternative class of tests is based on the discrepancy between the criterion function minimized with and without the restriction.

Criterion-based testing applies when we have a criterion function, say $J(\boldsymbol{\beta})$ with $\boldsymbol{\beta} \in \boldsymbol{B}$, which is minimized for estimation, and the goal is to test $\mathbb{H}_0 : \boldsymbol{\beta} \in \boldsymbol{B}_0$ versus $\mathbb{H}_1 : \boldsymbol{\beta} \notin \boldsymbol{B}_0$ where $\boldsymbol{B}_0 \subset \boldsymbol{B}$. Minimizing the criterion function over $\boldsymbol{B}$ and $\boldsymbol{B}_0$ we obtain the unrestricted and restricted estimators

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}\in\boldsymbol{B}}{\operatorname{argmin}}\ J(\boldsymbol{\beta}) \\ \widetilde{\boldsymbol{\beta}} &= \underset{\boldsymbol{\beta}\in\boldsymbol{B}_0}{\operatorname{argmin}}\ J(\boldsymbol{\beta}).\end{aligned}$$

The **criterion-based statistic** for $\mathbb{H}_0$ versus $\mathbb{H}_1$ is proportional to

$$\begin{aligned}J &= \min_{\boldsymbol{\beta}\in\boldsymbol{B}_0} J(\boldsymbol{\beta}) - \min_{\boldsymbol{\beta}\in\boldsymbol{B}} J(\boldsymbol{\beta}) \\ &= J(\widetilde{\boldsymbol{\beta}}) - J(\widehat{\boldsymbol{\beta}}).\end{aligned}$$

The criterion-based statistic J is sometimes called a **distance** statistic, a **minimum-distance** statistic, or a **likelihood-ratio-like** statistic.

Since $\boldsymbol{B}_0$ is a subset of $\boldsymbol{B}$, $J(\widetilde{\boldsymbol{\beta}}) \geq J(\widehat{\boldsymbol{\beta}})$ and thus $J \geq 0$. The statistic J measures the cost (on the criterion) of imposing the null restriction $\boldsymbol{\beta} \in \boldsymbol{B}_0$.

9.12 Minimum Distance Tests

The minimum distance test is a criterion-based test where $J(\boldsymbol{\beta})$ is the minimum distance criterion (8.20)

$$J(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \widehat{\boldsymbol{W}} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \tag{9.14}$$

with $\widehat{\boldsymbol{\beta}}$ the unrestricted (LS) estimator. The restricted estimator $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ minimizes (9.14) subject to $\boldsymbol{\beta} \in \boldsymbol{B}_0$. Observing that $J(\widehat{\boldsymbol{\beta}}) = 0$, the minimum distance statistic simplifies to

$$J = J(\widetilde{\boldsymbol{\beta}}_{\text{md}}) = n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{md}}\right)' \widehat{\boldsymbol{W}} \left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{md}}\right). \tag{9.15}$$

The efficient minimum distance estimator $\widetilde{\boldsymbol{\beta}}_{\text{emd}}$ is obtained by setting $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}$ in (9.14) and (9.15). The efficient minimum distance statistic for $\mathbb{H}_0 : \boldsymbol{\beta} \in \boldsymbol{B}_0$ is therefore

$$J^* = n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1} \left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right). \tag{9.16}$$

Consider the class of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$. In this case we know from (8.28) that the efficient minimum distance estimator $\widetilde{\boldsymbol{\beta}}_{\text{emd}}$ subject to the constraint $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$ is

$$\widetilde{\boldsymbol{\beta}}_{\text{emd}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)$$

and thus

$$\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}} = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right).$$

Substituting into (9.16) we find

$$\begin{aligned} J^* &= n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1} \boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right) \\ &= n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right) \\ &= W, \end{aligned} \tag{9.17}$$

which is the Wald statistic (9.10).

Thus for linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$, the efficient minimum distance statistic J^* is identical to the Wald statistic (9.10). For non-linear hypotheses, however, the Wald and minimum distance statistics are different.

Newey and West (1987) established the asymptotic null distribution of J^* for linear and non-linear hypotheses.

Theorem 9.12.1 *Under Assumptions 7.1.2 and 7.10.1, and* $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, *then* $J^* \xrightarrow{d} \chi^2_q$.

Testing using the minimum distance statistic J^* is similar to testing using the Wald statistic W. Critical values and p-values are computed using the χ^2_q distribution. $\mathbb{H}_0$ is rejected in favor of $\mathbb{H}_1$ if J^* exceeds the level α critical value, which can be calculated in MATLAB as `chi2inv(1-α,q)`. The asymptotic p-value is $p = 1 - G_q(J^*)$. In MATLAB, use the command `1-chi2cdf(J,q)`.

9.13 Minimum Distance Tests Under Homoskedasticity

If we set $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}/s^2$ in (9.14) we obtain the criterion (8.22)

$$J^0(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)/s^2.$$

A minimum distance statistic for $\mathbb{H}_0 : \boldsymbol{\beta} \in \boldsymbol{B}_0$ is

$$J^0 = \min_{\boldsymbol{\beta}\in\boldsymbol{B}_0} J^0(\boldsymbol{\beta}).$$

Equation (8.23) showed that

$$SSE(\boldsymbol{\beta}) = n\widehat{\sigma}^2 + s^2 J^0(\boldsymbol{\beta})$$

and so the minimizers of $SSE(\boldsymbol{\beta})$ and $J^0(\boldsymbol{\beta})$ are identical. Thus the constrained minimizer of $J^0(\boldsymbol{\beta})$ is constrained least-squares

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \operatorname*{argmin}_{\boldsymbol{\beta}\in\boldsymbol{B}_0} J^0(\boldsymbol{\beta}) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\boldsymbol{B}_0} SSE(\boldsymbol{\beta}) \tag{9.18}$$

and therefore

$$\begin{aligned} J_n^0 &= J_n^0(\widetilde{\boldsymbol{\beta}}_{\text{cls}}) \\ &= n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)' \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)/s^2. \end{aligned}$$

In the special case of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$, the constrained least-squares estimator subject to $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$ has the solution (8.10)

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)$$

and solving we find

$$J^0 = n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)'\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)/s^2 = W^0. \tag{9.19}$$

This is the homoskedastic Wald statistic (9.13). Thus for testing linear hypotheses, homoskedastic minimum distance and Wald statistics agree.

For nonlinear hypotheses they disagree, but have the same null asymptotic distribution.

Theorem 9.13.1 *Under Assumptions 7.1.2 and 7.10.1,* $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, *and* $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, *then* $J^0 \xrightarrow{d} \chi_q^2$.

9.14 F Tests

In Section 5.15 we introduced the F test for exclusion restrictions in the normal regression model. More generally, the F statistic for testing $\mathbb{H}_0 : \boldsymbol{\beta} \in \boldsymbol{B}_0$ is

$$F = \frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)/q}{\widehat{\sigma}^2/(n-k)} \tag{9.20}$$

where

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right)^2$$

and $\widehat{\boldsymbol{\beta}}$ are the unconstrained estimators of $\boldsymbol{\beta}$ and σ^2,

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)^2$$

and $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ are the constrained least-squares estimators from (9.18), q is the number of restrictions, and k is the number of unconstrained coefficients.

We can alternatively write

$$F = \frac{SSE(\widetilde{\boldsymbol{\beta}}_{\text{cls}}) - SSE(\widehat{\boldsymbol{\beta}})}{qs^2} \tag{9.21}$$

where

$$SSE(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2$$

is the sum-of-squared errors. Thus F is a criterion-based statistic. Using (8.23) we can also write F as

$$F = J^0/q,$$

so the F statistic is identical to the homoskedastic minimum distance statistic divided by the number of restrictions q.

As we discussed in the previous section, in the special case of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$, $J^0 = W^0$. It follows that in this case $F = W^0/q$. Thus for linear restrictions the F statistic equals the homoskedastic Wald statistic divided by q. It follows that they are equivalent tests for $\mathbb{H}_0$ against $\mathbb{H}_1$.

Theorem 9.14.1 *For tests of linear hypotheses* $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$,

$$F = W^0/q$$

the F statistic equals the homoskedastic Wald statistic divided by the degrees of freedom. Thus under 7.1.2 and 7.10.1, $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, *and* $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, *then*

$$F \xrightarrow{d} \chi_q^2/q.$$

When using an F statistic, it is conventional to use the $F_{q,n-k}$ distribution for critical values and p-values. Critical values are given in MATLAB by `finv(1-α,q,n-k)`, and p-values by `1-fcdf(F,q,n-k)`. Alternatively, the χ_q^2/q distribution can be used, using `chi2inv(1-α,q)/q` and `1-chi2cdf(F*q,q)`, respectively. Using the $F_{q,n-k}$ distribution is a prudent small sample adjustment which yields exact answers if the errors are normal, and otherwise slightly increasing the critical values and p-values relative to the asymptotic approximation. Once again, if the sample size is small enough that the choice makes a difference, then probably we shouldn't be trusting the asymptotic approximation anyway!

An elegant feature about (9.20) or (9.21) is that they are directly computable from the standard output from two simple OLS regressions, as the sum of squared errors (or regression variance) is a typical printed output from statistical packages, and is often reported in applied tables. Thus F can be calculated by hand from standard reported statistics even if you don't have the original data (or if you are sitting in a seminar and listening to a presentation!).

If you are presented with an F statistic (or a Wald statistic, as you can just divide by q) but don't have access to critical values, a useful rule of thumb is to know that for large n, the 5% asymptotic critical value is decreasing as q increases, and is less than 2 for $q \geq 7$.

A word of warning: In many statistical packages, when an OLS regression is estimated an "F-statistic" is automatically reported, even though no hypothesis test was requested. What the package is reporting is an F statistic of the hypothesis that all slope coefficients[1] are zero. This was a popular statistic in the early days of econometric reporting when sample sizes were very small and researchers wanted to know if there was "any explanatory power" to their regression. This is rarely an issue today, as sample sizes are typically sufficiently large that this F statistic is nearly always highly significant. While there are special cases where this F statistic is useful, these cases are not typical. As a general rule, there is no reason to report this F statistic.

9.15 Hausman Tests

Hausman (1978) introduced a general idea about how to test a hypothesis $\mathbb{H}_0$. If you have two estimators, one which is efficient under $\mathbb{H}_0$ but inconsistent under $\mathbb{H}_1$, and another which is consistent under $\mathbb{H}_1$, then construct a test as a quadratic form in the differences of the estimators. In the case of testing a hypothesis $\mathbb{H}_0 : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ let $\widehat{\boldsymbol{\beta}}_{\text{ols}}$ denote the unconstrained least-squares estimator and let $\widetilde{\boldsymbol{\beta}}_{\text{emd}}$ denote the efficient minimum distance estimator which imposes $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$. Both estimators are consistent under $\mathbb{H}_0$, but $\widetilde{\boldsymbol{\beta}}_{\text{emd}}$ is asymptotically efficient. Under $\mathbb{H}_1$, $\widehat{\boldsymbol{\beta}}_{\text{ols}}$ is consistent for $\boldsymbol{\beta}$ but $\widetilde{\boldsymbol{\beta}}_{\text{emd}}$ is inconsistent. The difference has the asymptotic distribution

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\right).$$

[1] All coefficients except the intercept.

Let $\boldsymbol{A}^{-}$ denote the Moore-Penrose generalized inverse. The Hausman statistic for $\mathbb{H}_0$ is

$$\begin{aligned} H &= \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)' \widehat{\text{avar}} \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)^{-} \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right) \\ &= n \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)' \left(\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}} \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\right)^{-} \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right). \end{aligned}$$

The matrix $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2} \widehat{\boldsymbol{R}} \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2}$ idempotent so its generalized inverse is itself. (See Section **??**.) It follows that

$$\begin{aligned} \left(\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}} \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\right)^{-} &= \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1/2} \left(\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2} \widehat{\boldsymbol{R}} \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2}\right)^{-} \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1/2} \\ &= \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1/2} \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2} \widehat{\boldsymbol{R}} \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{1/2} \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1/2} \\ &= \widehat{\boldsymbol{R}} \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}'. \end{aligned}$$

Thus the Hausman statistic is

$$H = n \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)' \widehat{\boldsymbol{R}} \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}' \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right),$$

In the context of linear restrictions, $\widehat{\boldsymbol{R}} = \boldsymbol{R}$ and $\boldsymbol{R}'\widetilde{\boldsymbol{\beta}} = \boldsymbol{\theta}_0$ so the statistic takes the form

$$H = n \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{R}} \left(\boldsymbol{R}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{\theta}_0\right),$$

which is precisely the Wald statistic. With nonlinear restrictions then can differ.

In either case we see that that the asymptotic null distribution of the Hausman statistic H is χ^2_q, so the appropriate test is to reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $H > c$ where c is a critical value taken from the χ^2_q distribution.

Theorem 9.15.1 *For general hypotheses the Hausman test statistic is*

$$H = n \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)' \widehat{\boldsymbol{R}} \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}' \left(\widehat{\boldsymbol{\beta}}_{\text{ols}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right),$$

and has the asymptotic distribution under $\mathbb{H}_0 : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$,

$$H \xrightarrow{d} \chi^2_q.$$

Jerry Hausman

Jerry Hausman (1946-) of the United States is a leading micro-econometrician, best known for his influential contributions on specification testing and panel data.

9.16 Score Tests

Score tests are traditionally derived in likelihood analysis, but can more generally be constructed from first-order conditions evaluated at restricted estimates. We focus on the likelihood derivation.

Given the log likelihood function $\log L(\boldsymbol{\beta}, \sigma^2)$, a restriction $\mathbb{H}_0 : \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$, and restricted estimators $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}^2$, the **score statistic** for $\mathbb{H}_0$ is defined as

$$S = \left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)\right)' \left(-\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)\right)^{-1} \left(\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2)\right).$$

The idea is that if the restriction is true, then the restricted estimators should be close to the maximum of the log-likelihood where the derivative should be small. However if the restriction is false then the restricted estimators should be distant from the maximum and the derivative should be large. Hence small values of S are expected under $\mathbb{H}_0$ and large values under $\mathbb{H}_1$. Tests of $\mathbb{H}_0$ thus reject for large values of S.

We explore the score statistic in the context of the normal regression model and linear hypotheses $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta}$. Recall that in the normal regression log-likelihood function is

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2.$$

The constrained MLE under linear hypotheses is constrained least squares

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R} \left[\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right]^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right) \\
\widetilde{e}_i &= y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}} \\
\widetilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} \widetilde{e}_i^2.
\end{aligned}$$

We can calculate that the derivative and Hessian are

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) &= \frac{1}{\widetilde{\sigma}^2} \sum_{i=1}^{n} \boldsymbol{x}_i \left(y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}\right) = \frac{1}{\widetilde{\sigma}^2} \boldsymbol{X}'\widetilde{\boldsymbol{e}} \\
-\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) &= \frac{1}{\widetilde{\sigma}^2} \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' = \frac{1}{\widetilde{\sigma}^2} \boldsymbol{X}'\boldsymbol{X}.
\end{aligned}$$

Since $\widetilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$ we can further calculate that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \log L(\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}^2) &= \frac{1}{\widetilde{\sigma}^2} \left(\boldsymbol{X}'\boldsymbol{X}\right) \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right) \\
&= \frac{1}{\widetilde{\sigma}^2} \left(\boldsymbol{X}'\boldsymbol{X}\right) \left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}\right) \\
&= \frac{1}{\widetilde{\sigma}^2} \boldsymbol{R} \left[\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right]^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right).
\end{aligned}$$

Together we find that

$$S = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)' \left(\boldsymbol{R}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right) / \widetilde{\sigma}^2$$

This is identical to the homoskedastic Wald statistic, with s^2 replaced by $\widetilde{\sigma}^2$. We can also write S as a monotonic transformation of the F statistic, since

$$S = n \frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)}{\widetilde{\sigma}^2} = n \left(1 - \frac{\widehat{\sigma}^2}{\widetilde{\sigma}^2}\right) = n \left(1 - \frac{1}{1 + \frac{q}{n-k} F}\right).$$

The test "Reject $\mathbb{H}_0$ for large values of S" is identical to the test "Reject $\mathbb{H}_0$ for large values of F", so they are identical tests. Since for the normal regression model the exact distribution of F is known, it is better to use the F statistic with F p-values.

In more complicated settings a potential advantage of score tests is that they are calculated using the restricted parameter estimates $\widetilde{\boldsymbol{\beta}}$ rather than the unrestricted estimates $\widehat{\boldsymbol{\beta}}$. Thus when $\widetilde{\boldsymbol{\beta}}$ is relatively easy to calculate there can be a preference for score statistics. This is not a concern for linear restrictions.

More generally, score and score-like statistics can be constructed from first-order conditions evaluated at restricted parameter estimates. Also, when test statistics are constructed using covariance matrix estimators which are calculated using restricted parameter estimates (e.g. restricted residuals) then these are often described as score tests.

An example of the latter is the Wald-type statistic

$$W = \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{R}}' \widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}\right)^{-1} \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)$$

where the covariance matrix estimate $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is calculated using the restricted residuals $\widetilde{e}_i = y_i - \boldsymbol{x}_i' \widetilde{\boldsymbol{\beta}}$. This may be done when $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are high-dimensional, so there is wory that the estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is imprecise.

9.17 Problems with Tests of Nonlinear Hypotheses

While the t and Wald tests work well when the hypothesis is a linear restriction on $\boldsymbol{\beta}$, they can work quite poorly when the restrictions are nonlinear. This can be seen by a simple example introduced by Lafontaine and White (1986). Take the model

$$\begin{aligned} y_i &= \beta + e_i \\ e_i &\sim \mathrm{N}(0, \sigma^2) \end{aligned}$$

and consider the hypothesis

$$\mathbb{H}_0 : \beta = 1.$$

Let $\widehat{\beta}$ and $\widehat{\sigma}^2$ be the sample mean and variance of y_i. The standard Wald test for $\mathbb{H}_0$ is

$$W = n \frac{\left(\widehat{\beta} - 1\right)^2}{\widehat{\sigma}^2}.$$

Now notice that $\mathbb{H}_0$ is equivalent to the hypothesis

$$\mathbb{H}_0(s) : \beta^s = 1$$

for any positive integer s. Letting $r(\beta) = \beta^s$, and noting $\boldsymbol{R} = s\beta^{s-1}$, we find that the standard Wald test for $\mathbb{H}_0(s)$ is

$$W(s) = n \frac{\left(\widehat{\beta}^s - 1\right)^2}{\widehat{\sigma}^2 s^2 \widehat{\beta}^{2s-2}}.$$

While the hypothesis $\beta^s = 1$ is unaffected by the choice of s, the statistic $W(s)$ varies with s. This is an unfortunate feature of the Wald statistic.

To demonstrate this effect, we have plotted in Figure 9.1 the Wald statistic $W(s)$ as a function of s, setting $n/\widehat{\sigma}^2 = 10$. The increasing solid line is for the case $\widehat{\beta} = 0.8$. The decreasing dashed line is for the case $\widehat{\beta} = 1.6$. It is easy to see that in each case there are values of s for which the test statistic is significant relative to asymptotic critical values, while there are other values of s

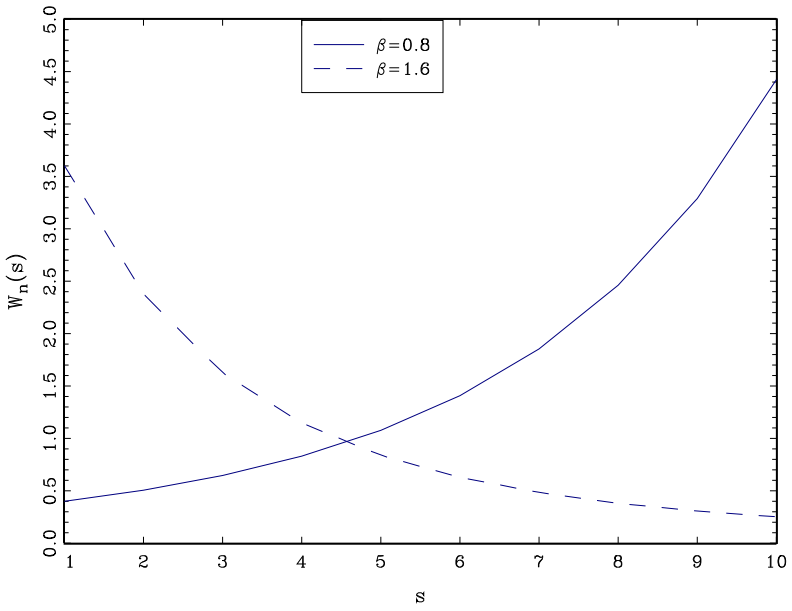


Figure 9.1: Wald Statistic as a function of s

for which the test statistic is insignificant. This is distressing since the choice of s is arbitrary and irrelevant to the actual hypothesis.

Our first-order asymptotic theory is not useful to help pick s, as $W(s) \xrightarrow{d} \chi^2_1$ under $\mathbb{H}_0$ for any s. This is a context where **Monte Carlo simulation** can be quite useful as a tool to study and compare the exact distributions of statistical procedures in finite samples. The method uses random simulation to create artificial datasets, to which we apply the statistical tools of interest. This produces random draws from the statistic's sampling distribution. Through repetition, features of this distribution can be calculated.

In the present context of the Wald statistic, one feature of importance is the Type I error of the test using the asymptotic 5% critical value 3.84 – the probability of a false rejection, $\Pr(W(s) > 3.84 \mid \beta = 1)$. Given the simplicity of the model, this probability depends only on s, n, and σ^2. In Table 9.1 we report the results of a Monte Carlo simulation where we vary these three parameters. The value of s is varied from 1 to 10, n is varied among 20, 100 and 500, and σ is varied among 1 and 3. The Table reports the simulation estimate of the Type I error probability from 50,000 random samples. Each row of the table corresponds to a different value of s – and thus corresponds to a particular choice of test statistic. The second through seventh columns contain the Type I error probabilities for different combinations of n and σ. These probabilities are calculated as the percentage of the 50,000 simulated Wald statistics $W(s)$ which are larger than 3.84. The null hypothesis $\beta^s = 1$ is true, so these probabilities are Type I error.

To interpret the table, remember that the ideal Type I error probability is 5% (.05) with deviations indicating distortion. Type I error rates between 3% and 8% are considered reasonable. Error rates above 10% are considered excessive. Rates above 20% are unacceptable. When comparing statistical procedures, we compare the rates row by row, looking for tests for which rejection rates are close to 5% and rarely fall outside of the 3%-8% range. For this particular example the only test which meets this criterion is the conventional $W = W(1)$ test. Any other choice of s leads to a test with unacceptable Type I error probabilities.

Table 9.1
Type I Error Probability of Asymptotic 5% $W(s)$ Test

	$\sigma = 1$			$\sigma = 3$		
s	$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$
1	.06	.05	.05	.07	.05	.05
2	.08	.06	.05	.15	.08	.06
3	.10	.06	.05	.21	.12	.07
4	.13	.07	.06	.25	.15	.08
5	.15	.08	.06	.28	.18	.10
6	.17	.09	.06	.30	.20	.11
7	.19	.10	.06	.31	.22	.13
8	.20	.12	.07	.33	.24	.14
9	.22	.13	.07	.34	.25	.15
10	.23	.14	.08	.35	.26	.16

Note: Rejection frequencies from 50,000 simulated random samples

In Table 9.1 you can also see the impact of variation in sample size. In each case, the Type I error probability improves towards 5% as the sample size n increases. There is, however, no magic choice of n for which all tests perform uniformly well. Test performance deteriorates as s increases, which is not surprising given the dependence of $W(s)$ on s as shown in Figure 9.1.

In this example it is not surprising that the choice $s = 1$ yields the best test statistic. Other choices are arbitrary and would not be used in practice. While this is clear in this particular example, in other examples natural choices are not always obvious and the best choices may in fact appear counter-intuitive at first.

This point can be illustrated through another example which is similar to one developed in Gregory and Veall (1985). Take the model

$$\begin{aligned} y_i &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned} \tag{9.22}$$

and the hypothesis

$$\mathbb{H}_0 : \frac{\beta_1}{\beta_2} = \theta_0$$

where θ_0 is a known constant. Equivalently, define $\theta = \beta_1/\beta_2$, so the hypothesis can be stated as $\mathbb{H}_0 : \theta = \theta_0$.

Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ be the least-squares estimates of (9.22), let $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ be an estimate of the covariance matrix for $\widehat{\boldsymbol{\beta}}$ and set $\widehat{\theta} = \widehat{\beta}_1/\widehat{\beta}_2$. Define

$$\widehat{\boldsymbol{R}}_1 = \begin{pmatrix} 0 \\ \dfrac{1}{\widehat{\beta}_2} \\ -\dfrac{\widehat{\beta}_1}{\widehat{\beta}_2^2} \end{pmatrix}$$

so that the standard error for $\widehat{\theta}$ is $s(\widehat{\theta}) = \left(\widehat{\boldsymbol{R}}_1' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}_1\right)^{1/2}$. In this case a t-statistic for $\mathbb{H}_0$ is

$$T_1 = \frac{\left(\dfrac{\widehat{\beta}_1}{\widehat{\beta}_2} - \theta_0\right)}{s(\widehat{\theta})}.$$

An alternative statistic can be constructed through reformulating the null hypothesis as

$$\mathbb{H}_0 : \beta_1 - \theta_0 \beta_2 = 0.$$

A t-statistic based on this formulation of the hypothesis is

$$T_2 = \frac{\widehat{\beta}_1 - \theta_0 \widehat{\beta}_2}{\left(\boldsymbol{R}_2' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \boldsymbol{R}_2 \right)^{1/2}}.$$

where

$$\boldsymbol{R}_2 = \begin{pmatrix} 0 \\ 1 \\ -\theta_0 \end{pmatrix}.$$

To compare T_1 and T_2 we perform another simple Monte Carlo simulation. We let x_{1i} and x_{2i} be mutually independent $\mathrm{N}(0,1)$ variables, e_i be an independent $\mathrm{N}(0,\sigma^2)$ draw with $\sigma = 3$, and normalize $\beta_0 = 0$ and $\beta_1 = 1$. This leaves β_2 as a free parameter, along with sample size n. We vary β_2 among .1, .25, .50, .75, and 1.0 and n among 100 and 500.

Table 9.2
Type I Error Probability of Asymptotic 5% t-tests

	$n = 100$				$n = 500$			
	$\Pr(T < -1.645)$		$\Pr(T > 1.645)$		$\Pr(T < -1.645)$		$\Pr(T > 1.645)$	
β_2	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
.10	.47	.06	.00	.06	.28	.05	.00	.05
.25	.26	.06	.00	.06	.15	.05	.00	.05
.50	.15	.06	.00	.06	.10	.05	.00	.05
.75	.12	.06	.00	.06	.09	.05	.00	.05
1.00	.10	.06	.00	.06	.07	.05	.02	.05

The one-sided Type I error probabilities $\Pr(T < -1.645)$ and $\Pr(T > 1.645)$ are calculated from 50,000 simulated samples. The results are presented in Table 9.2. Ideally, the entries in the table should be 0.05. However, the rejection rates for the T_1 statistic diverge greatly from this value, especially for small values of β_2. The left tail probabilities $\Pr(T_1 < -1.645)$ greatly exceed 5%, while the right tail probabilities $\Pr(T_1 > 1.645)$ are close to zero in most cases. In contrast, the rejection rates for the linear T_2 statistic are invariant to the value of β_2, and are close to the ideal 5% rate for both sample sizes. The implication of Table 8.2 is that the two t-ratios have dramatically different sampling behavior.

The common message from both examples is that Wald statistics are sensitive to the algebraic formulation of the null hypothesis.

A simple solution is to use the minimum distance statistic J, which equals W with $r = 1$ in the first example, and $|T_2|$ in the second example. The minimum distance statistic is invariant to the algebraic formulation of the null hypothesis, so is immune to this problem. Whenever possible, the Wald statistic should not be used to test nonlinear hypotheses.

9.18 Monte Carlo Simulation

In Section 9.17 we introduced the method of Monte Carlo simulation to illustrate the small sample problems with tests of nonlinear hypotheses. In this section we describe the method in more detail.

Recall, our data consist of observations $(y_i, \boldsymbol{x}_i)$ which are random draws from a population distribution F. Let $\boldsymbol{\theta}$ be a parameter and let $T = T\left((y_1, \boldsymbol{x}_1), ..., (y_n, \boldsymbol{x}_n), \boldsymbol{\theta}\right)$ be a statistic of interest, for example an estimator $\widehat{\theta}$ or a t-statistic $(\widehat{\theta} - \theta)/s(\widehat{\theta})$. The exact distribution of T is

$$G(u, F) = \Pr(T \leq u \mid F).$$

While the asymptotic distribution of T might be known, the exact (finite sample) distribution G is generally unknown.

Monte Carlo simulation uses numerical simulation to compute $G(u, F)$ for selected choices of F. This is useful to investigate the performance of the statistic T in reasonable situations and sample sizes. The basic idea is that for any given F, the distribution function $G(u, F)$ can be calculated numerically through simulation. The name Monte Carlo derives from the famous Mediterranean gambling resort where games of chance are played.

The method of Monte Carlo is quite simple to describe. The researcher chooses F (the distribution of the data) and the sample size n. A "true" value of $\boldsymbol{\theta}$ is implied by this choice, or equivalently the value $\boldsymbol{\theta}$ is selected directly by the researcher which implies restrictions on F.

Then the following experiment is conducted by computer simulation:

1. n independent random pairs $(y_i^*, \boldsymbol{x}_i^*)$, $i = 1, ..., n$, are drawn from the distribution F using the computer's random number generator.

2. The statistic $T = T\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), \boldsymbol{\theta}\right)$ is calculated on this pseudo data.

For step 1, computer packages have built-in random number procedures including U[0, 1] and N(0, 1). From these most random variables can be constructed. (For example, a chi-square can be generated by sums of squares of normals.)

For step 2, it is important that the statistic be evaluated at the "true" value of $\boldsymbol{\theta}$ corresponding to the choice of F.

The above experiment creates one random draw from the distribution $G(u, F)$. This is one observation from an unknown distribution. Clearly, from one observation very little can be said. So the researcher repeats the experiment B times, where B is a large number. Typically, we set $B = 1000$ or $B = 5000$. We will discuss this choice later.

Notationally, let the b^{th} experiment result in the draw T_b, $b = 1, ..., B$. These results are stored. After all B experiments have been calculated, these results constitute a random sample of size B from the distribution of $G(u, F) = \Pr(T_b \leq u) = \Pr(T \leq u \mid F)$.

From a random sample, we can estimate any feature of interest using (typically) a method of moments estimator. We now describe some specific examples.

Suppose we are interested in the bias, mean-squared error (MSE), and/or variance of the distribution of $\widehat{\theta} - \theta$. We then set $T = \widehat{\theta} - \theta$, run the above experiment, and calculate

$$\widehat{Bias(\hat{\theta})} = \frac{1}{B}\sum_{b=1}^{B} T_b = \frac{1}{B}\sum_{b=1}^{B} \widehat{\theta}_b - \theta$$

$$\widehat{MSE(\hat{\theta})} = \frac{1}{B}\sum_{b=1}^{B} (T_b)^2 = \frac{1}{B}\sum_{b=1}^{B} \left(\widehat{\theta}_b - \theta\right)^2$$

$$\widehat{\text{var}(\widehat{\theta})} = \widehat{MSE(\widehat{\theta})} - \left(\widehat{Bias(\widehat{\theta})}\right)^2$$

Suppose we are interested in the Type I error associated with an asymptotic 5% two-sided t-test. We would then set $T = \left|\widehat{\theta} - \theta\right| / s(\widehat{\theta})$ and calculate

$$\widehat{P} = \frac{1}{B}\sum_{b=1}^{B} 1\left(T_b \geq 1.96\right), \tag{9.23}$$

the percentage of the simulated t-ratios which exceed the asymptotic 5% critical value.

Suppose we are interested in the 5% and 95% quantile of $T = \widehat{\theta}$ or $T = \left(\widehat{\theta} - \theta\right) / s(\widehat{\theta})$. We then compute the 5% and 95% sample quantiles of the sample $\{T_b\}$. The α sample quantile is a number

q_α such that $100\alpha\%$ of the sample are less than q_α. A simple way to compute sample quantiles is to sort the sample $\{T_b\}$ from low to high. Then q_α is the N^{th} number in this ordered sequence, where $N = (B+1)\alpha$. It is therefore convenient to pick B so that N is an integer. For example, if we set $B = 999$, then the 5% sample quantile is 50^{th} sorted value and the 95% sample quantile is the 950^{th} sorted value.

The typical purpose of a Monte Carlo simulation is to investigate the performance of a statistical procedure (estimator or test) in realistic settings. Generally, the performance will depend on n and F. In many cases, an estimator or test may perform wonderfully for some values, and poorly for others. It is therefore useful to conduct a variety of experiments, for a selection of choices of n and F.

As discussed above, the researcher must select the number of experiments, B. Often this is called the number of **replications**. Quite simply, a larger B results in more precise estimates of the features of interest of G, but requires more computational time. In practice, therefore, the choice of B is often guided by the computational demands of the statistical procedure. Since the results of a Monte Carlo experiment are estimates computed from a random sample of size B, it is straightforward to calculate standard errors for any quantity of interest. If the standard error is too large to make a reliable inference, then B will have to be increased.

In particular, it is simple to make inferences about rejection probabilities from statistical tests, such as the percentage estimate reported in (9.23). The random variable $1(T_b \geq 1.96)$ is iid Bernoulli, equalling 1 with probability $p = \mathbb{E}(1(T_b \geq 1.96))$. The average (9.23) is therefore an unbiased estimator of p with standard error $s(\widehat{p}) = \sqrt{p(1-p)/B}$. As p is unknown, this may be approximated by replacing p with $\widehat{p}$ or with an hypothesized value. For example, if we are assessing an asymptotic 5% test, then we can set $s(\widehat{p}) = \sqrt{(.05)(.95)/B} \simeq .22/\sqrt{B}$. Hence, standard errors for $B = 100$, 1000, and 5000, are, respectively, $s(\widehat{p}) = .022$, .007, and .003.

Most papers in econometric methods, and some empirical papers, include the results of Monte Carlo simulations to illustrate the performance of their methods. When extending existing results, it is good practice to start by replicating existing (published) results. This is not exactly possible in the case of simulation results, as they are inherently random. For example suppose a paper investigates a statistical test, and reports a simulated rejection probability of 0.07 based on a simulation with $B = 100$ replications. Suppose you attempt to replicate this result, and find a rejection probability of 0.03 (again using $B = 100$ simulation replications). Should you conclude that you have failed in your attempt? Absolutely not! Under the hypothesis that both simulations are identical, you have two independent estimates, $\widehat{p}_1 = 0.07$ and $\widehat{p}_2 = 0.03$, of a common probability p. The asymptotic (as $B \to \infty$) distribution of their difference is $\sqrt{B}(\widehat{p}_1 - \widehat{p}_2) \xrightarrow{d} \mathrm{N}(0, 2p(1-p))$, so a standard error for $\widehat{p}_1 - \widehat{p}_2 = 0.04$ is $\widehat{s} = \sqrt{2p(1-p)/B} \simeq 0.03$, using the estimate $p = (\widehat{p}_1 + \widehat{p}_2)/2$. Since the t-ratio $0.04/0.03 = 1.3$ is not statistically significant, it is incorrect to reject the null hypothesis that the two simulations are identical. The difference between the results $\widehat{p}_1 = 0.07$ and $\widehat{p}_2 = 0.03$ is consistent with random variation.

What should be done? The first mistake was to copy the previous paper's choice of $B = 100$. Instead, suppose you set $B = 5000$. Suppose you now obtain $\widehat{p}_2 = 0.04$. Then $\widehat{p}_1 - \widehat{p}_2 = 0.03$ and a standard error is $\widehat{s} = \sqrt{p(1-p)(1/100 + 1/5000)} \simeq 0.02$. Still we cannot reject the hypothesis that the two simulations are different. Even though the estimates (0.07 and 0.04) appear to be quite different, the difficulty is that the original simulation used a very small number of replications ($B = 100$) so the reported estimate is quite imprecise. In this case, it is appropriate to conclude that your results "replicate" the previous study, as there is no statistical evidence to reject the hypothesis that they are equivalent.

Most journals have policies requiring authors to make available their data sets and computer programs required for empirical results. They do not have similar policies regarding simulations. Nevertheless, it is good professional practice to make your simulations available. The best practice is to post your simulation code on your webpage. This invites others to build on and use your results, leading to possible collaboration, citation, and/or advancement.

9.19 Confidence Intervals by Test Inversion

There is a close relationship between hypothesis tests and confidence intervals. We observed in Section 7.13 that the standard 95% asymptotic confidence interval for a parameter θ is

$$\begin{aligned}\widehat{C} &= \left[\widehat{\theta} - 1.96 \cdot s(\widehat{\theta}),\quad \widehat{\theta} + 1.96 \cdot s(\widehat{\theta})\right] \\ &= \{\theta : |T(\theta)| \leq 1.96\}.\end{aligned} \tag{9.24}$$

That is, we can describe $\widehat{C}$ as "The point estimate plus or minus 2 standard errors" or "The set of parameter values not rejected by a two-sided t-test." The second definition, known as **test statistic inversion** is a general method for finding confidence intervals, and typically produces confidence intervals with excellent properties.

Given a test statistic $T(\theta)$ and critical value c, the acceptance region "Accept if $T(\theta) \leq c$" is identical to the confidence interval $\widehat{C} = \{\theta : T(\theta) \leq c\}$. Since the regions are identical, the probability of coverage $\Pr\left(\theta \in \widehat{C}\right)$ equals the probability of correct acceptance $\Pr(\text{Accept}|\theta)$ which is exactly 1 minus the Type I error probability. Thus inverting a test with good Type I error probabilities yields a confidence interval with good coverage probabilities.

Now suppose that the parameter of interest $\theta = r(\boldsymbol{\beta})$ is a nonlinear function of the coefficient vector $\boldsymbol{\beta}$. In this case the standard confidence interval for θ is the set $\widehat{C}$ as in (9.24) where $\widehat{\theta} = r(\widehat{\boldsymbol{\beta}})$ is the point estimate and $s(\widehat{\theta}) = \sqrt{\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}}$ is the delta method standard error. This confidence interval is inverting the t-test based on the nonlinear hypothesis $r(\boldsymbol{\beta}) = \theta$. The trouble is that in Section 9.17 we learned that there is no unique t-statistic for tests of nonlinear hypotheses and that the choice of parameterization matters greatly.

For example, if $\theta = \beta_1/\beta_2$ then the coverage probability of the standard interval (9.24) is 1 minus the probability of the Type I error, which as shown in Table 8.2 can be far from the nominal 5%.

In this example a good solution is the same as discussed in Section 9.17 – to rewrite the hypothesis as a linear restriction. The hypothesis $\theta = \beta_1/\beta_2$ is the same as $\theta\beta_2 = \beta_1$. The t-statistic for this restriction is

$$T(\theta) = \frac{\widehat{\beta}_1 - \widehat{\beta}_2\theta}{\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{R}\right)^{1/2}}$$

where

$$\boldsymbol{R} = \begin{pmatrix} 1 \\ -\theta \end{pmatrix}$$

and $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is the covariance matrix for $(\widehat{\beta}_1\ \widehat{\beta}_2)$. A 95% confidence interval for $\theta = \beta_1/\beta_2$ is the set of values of θ such that $|T(\theta)| \leq 1.96$. Since θ appears in both the numerator and denominator, $T(\theta)$ is a non-linear function of θ so the easiest method to find the confidence set is by grid search over θ.

For example, in the wage equation

$$\log(Wage) = \beta_1 Experience + \beta_2 Experience^2/100 + \cdots$$

the highest expected wage occurs at $Experience = -50\beta_1/\beta_2$. From Table 4.1 we have the point estimate $\widehat{\theta} = 29.8$ and we can calculate the standard error $s(\widehat{\theta}) = 0.022$ for a 95% confidence interval [29.8, 29.9]. However, if we instead invert the linear form of the test we can numerically find the interval [29.1, 30.6] which is much larger. From the evidence presented in Section 9.17 we know the first interval can be quite inaccurate and the second interval is greatly preferred.

9.20 Multiple Tests and Bonferroni Corrections

In most applications, economists examine a large number of estimates, test statistics, and p-values. What does it mean (or does it mean anything) if one statistic appears to be "significant" after examining a large number of statistics? This is known as the problem of **multiple testing** or **multiple comparisons**.

To be specific, suppose we examine a set of k coefficients, standard errors and t-ratios, and consider the "significance" of each statistic. Based on conventional reasoning, for each coefficient we would reject the hypothesis that the coefficient is zero with asymptotic size α if the absolute t-statistic exceeds the $1-\alpha$ critical value of the normal distribution, or equivalently if the p-value for the t-statistic is smaller than α. If we observe that one of the k statistics is "significant" based on this criteria, that means that one of the p-values is smaller than α, or equivalently, that the smallest p-value is smaller than α. We can then rephrase the question: Under the joint hypothesis that a set of k hypotheses are all true, what is the probability that the smallest p-value is smaller than α? In general, we cannot provide a precise answer to this quesion, but the Bonferroni correction bounds this probability by αk. The Bonferroni method furthermore suggests that if we want the familywise error probability (the probability that one of the tests falsely rejects) is bounded below α, then an appropriate rule is to reject only if the smallest p-value is smaller than α/k. Equivalenlty, the Bonferroni familywise p-value is $k \min_{j \leq k} p_j$.

Formally, suppose we have k hypotheses $\mathbb{H}_j$, $j = 1, ..., k$. For each we have a test and associated p-value p_j with the property that when $\mathbb{H}_j$ is true $\lim_{n\to\infty} \Pr(p_j < \alpha) = \alpha$. We then observe that among the k tests, one of the k will appear "significant" if $\min_{j \leq k} p_j < \alpha$. This event can be written as

$$\left\{ \min_{j \leq k} p_j < \alpha \right\} = \bigcup_{j=1}^{k} \{p_j < \alpha\}.$$

Boole's inequality states that for any k events A_j, $\Pr\left(\bigcup_{j=1}^{k} A_j\right) \leq \sum_{j=1}^{k} \Pr(A_k)$. Thus

$$\Pr\left(\min_{j \leq k} p_j < \alpha \right) \leq \sum_{j=1}^{k} \Pr(p_j < \alpha) \longrightarrow k\alpha$$

as stated. This demonstates that the familywise rejection probability is at most k times the individual rejection probability.

Furthermore,

$$\Pr\left(\min_{j \leq k} p_j < \frac{\alpha}{k} \right) \leq \sum_{j=1}^{k} \Pr\left(p_j < \frac{\alpha}{k} \right) \longrightarrow \alpha.$$

This demonstrates that the family rejection probability can be controlled (bounded below α) if each individual test is subjected to the stricter standard that a p-value must be smaller than α/k to be labeled as "significant."

To illustrate, suppose we have two coefficient estimates, with individual p-values 0.04 and 0.15. Based on a conventional 5% level, the standard individual tests would suggest that the first coefficient estimate is "significant" but not the second. A Bonferroni 5% test, however, does not reject as it would require that the smallest p-value be smaller than 0.025, which is not the case in this example. Alternatively, the Bonferroni familywise p-value is 0.08, which is not significant at the 5% level.

In contrast, if the two p-values are 0.01 and 0.15, then the Bonferroni familywise p-value is 0.02, which is significant at the 5% level.

9.21 Power and Test Consistency

The **power** of a test is the probability of rejecting $\mathbb{H}_0$ when $\mathbb{H}_1$ is true.

For simplicity suppose that y_i is i.i.d. $N(\theta, \sigma^2)$ with σ^2 known, consider the t-statistic $T(\theta) = \sqrt{n}(\bar{y} - \theta)/\sigma$, and tests of $\mathbb{H}_0 : \theta = 0$ against $\mathbb{H}_1 : \theta > 0$. We reject $\mathbb{H}_0$ if $T = T(0) > c$. Note that

$$T = T(\theta) + \sqrt{n}\theta/\sigma$$

and $T(\theta)$ has an exact $\mathrm{N}(0,1)$ distribution. This is because $T(\theta)$ is centered at the true mean θ, while the test statistic $T(0)$ is centered at the (false) hypothesized mean of 0.

The power of the test is

$$\Pr(T > c \mid \theta) = \Pr\left(\mathrm{Z} + \sqrt{n}\theta/\sigma > c\right) = 1 - \Phi\left(c - \sqrt{n}\theta/\sigma\right).$$

This function is monotonically increasing in μ and n, and decreasing in σ and c.

Notice that for any c and $\theta \neq 0$, the power increases to 1 as $n \to \infty$. This means that for $\theta \in \mathbb{H}_1$, the test will reject $\mathbb{H}_0$ with probability approaching 1 as the sample size gets large. We call this property **test consistency**.

Definition 9.21.1 *A test of $\mathbb{H}_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ is **consistent against fixed alternatives** if for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, $\Pr(\text{Reject } \mathbb{H}_0 \mid \theta) \to 1$ as $n \to \infty$.*

For tests of the form "Reject $\mathbb{H}_0$ if $T > c$", a sufficient condition for test consistency is that the T diverges to positive infinity with probability one for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$.

Definition 9.21.2 *$T \xrightarrow{p} \infty$ as $n \to \infty$ if for all $M < \infty$, $\Pr(T \leq M) \to 0$ as $n \to \infty$. Similarly, $T \xrightarrow{p} -\infty$ as $n \to \infty$ if for all $M < \infty$, $\Pr(T \geq -M) \to 0$ as $n \to \infty$.*

In general, t-tests and Wald tests are consistent against fixed alternatives. Take a t-statistic for a test of $\mathbb{H}_0 : \theta = \theta_0$

$$T = \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})}$$

where θ_0 is a known value and $s(\widehat{\theta}) = \sqrt{n^{-1}\widehat{V}_\theta}$. Note that

$$T = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} + \frac{\sqrt{n}(\theta - \theta_0)}{\sqrt{\widehat{V}_\theta}}.$$

The first term on the right-hand-side converges in distribution to $\mathrm{N}(0,1)$. The second term on the right-hand-side equals zero if $\theta = \theta_0$, converges in probability to $+\infty$ if $\theta > \theta_0$, and converges in probability to $-\infty$ if $\theta < \theta_0$. Thus the two-sided t-test is consistent against $\mathbb{H}_1 : \theta \neq \theta_0$, and one-sided t-tests are consistent against the alternatives for which they are designed.

Theorem 9.21.1 *Under Assumptions 7.1.2 and 7.10.1, for $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$ and $q = 1$, then $|T| \xrightarrow{p} \infty$, so for any $c < \infty$ the test "Reject $\mathbb{H}_0$ if $|T| > c$" is consistent against fixed alternatives.*

The Wald statistic for $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ against $\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ is

$$W = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).$$

Under $\mathbb{H}_1$, $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. Thus $\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{p} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)' \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) > 0$. Hence under $\mathbb{H}_1$, $W \xrightarrow{p} \infty$. Again, this implies that Wald tests are consistent tests.

Theorem 9.21.2 *Under Assumptions 7.1.2 and 7.10.1, for $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$, then $W \xrightarrow{p} \infty$, so for any $c < \infty$ the test "Reject $\mathbb{H}_0$ if $W > c$" is consistent against fixed alternatives.*

9.22 Asymptotic Local Power

Consistency is a good property for a test, but does not give a useful approximation to the power of a test. To approximate the power function we need a distributional approximation.

The standard asymptotic method for power analysis uses what are called **local alternatives**. This is similar to our analysis of restriction estimation under misspecification (Section 8.13). The technique is to index the parameter by sample size so that the asymptotic distribution of the statistic is continuous in a localizing parameter. In this section we consider t-tests on real-valued parameters and in the next section consider Wald tests. Specifically, we consider parameter vectors $\boldsymbol{\beta}_n$ which are indexed by sample size n and satisfy the real-valued relationship

$$\theta_n = r(\boldsymbol{\beta}_n) = \theta_0 + n^{-1/2} h \tag{9.25}$$

where the scalar h is called a **localizing parameter**. We index $\boldsymbol{\beta}_n$ and θ_n by sample size to indicate their dependence on n. The way to think of (9.25) is that the true value of the parameters are $\boldsymbol{\beta}_n$ and θ_n. The parameter θ_n is close to the hypothesized value θ_0, with deviation $n^{-1/2}h$.

The specification (9.25) states that for any fixed h , θ_n approaches θ_0 as n gets large. Thus θ_n is "close" or "local" to θ_0. The concept of a localizing sequence (9.25) might seem odd since in the actual world the sample size cannot mechanically affect the value of the parameter. Thus (9.25) should not be interpreted literally. Instead, it should be interpreted as a technical device which allows the asymptotic distribution of the test statistic to be continuous in the alternative hypothesis.

To evaluate the asymptotic distribution of the test statistic we start by examining the scaled estimate centered at the hypothesized value θ_0. Breaking it into a term centered at the true value θ_n and a remainder we find

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta} - \theta_0\right) &= \sqrt{n}\left(\widehat{\theta} - \theta_n\right) + \sqrt{n}\left(\theta_n - \theta_0\right) \\ &= \sqrt{n}\left(\widehat{\theta} - \theta_n\right) + h\end{aligned}$$

where the second equality is (9.25). The first term is asymptotically normal:

$$\sqrt{n}\left(\widehat{\theta} - \theta_n\right) \xrightarrow{d} \sqrt{V_\theta}\mathrm{Z}.$$

where $\mathrm{Z} \sim \mathrm{N}(0, 1)$. Therefore

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \sqrt{V_\theta}\mathrm{Z} + h$$

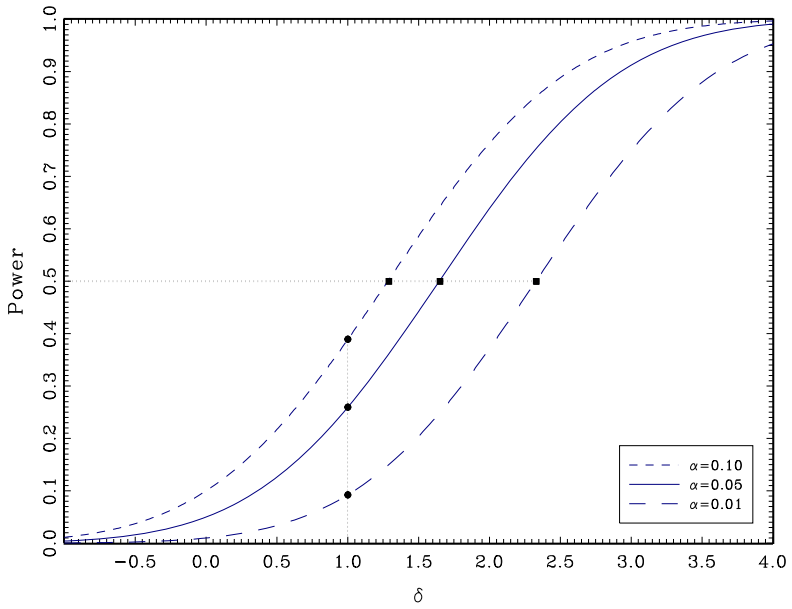


Figure 9.2: Asymptotic Local Power Function of One-Sided t Test

or $\mathrm{N}(h, V_\theta)$. This is a continuous asymptotic distribution, and depends continuously on the localizing parameter h.

Applied to the t statistic we find

$$\begin{aligned} T &= \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})} \\ &\xrightarrow{d} \frac{\sqrt{V_\theta}\mathrm{Z} + h}{\sqrt{V_\theta}} \\ &\sim \mathrm{Z} + \delta \end{aligned} \tag{9.26}$$

where $\delta = h/\sqrt{V_\theta}$. This generalizes Theorem 9.4.1 (which assumes $\mathbb{H}_0$ is true) to allow for local alternatives of the form (9.25).

Consider a t-test of $\mathbb{H}_0$ against the one-sided alternative $\mathbb{H}_1 : \theta > \theta_0$ which rejects $\mathbb{H}_0$ for $T > c$ where $\Phi(c) = 1 - \alpha$. The **asymptotic local power** of this test is the limit (as the sample size diverges) of the rejection probability under the local alternative (9.25)

$$\begin{aligned} \lim_{n\to\infty} \Pr\left(\text{Reject } \mathbb{H}_0\right) &= \lim_{n\to\infty} \Pr\left(T > c\right) \\ &= \Pr\left(\mathrm{Z} + \delta > c\right) \\ &= 1 - \Phi\left(c - \delta\right) \\ &= \Phi\left(\delta - c\right) \\ &\overset{def}{=} \pi(\delta). \end{aligned}$$

We call $\pi(\delta)$ the **asymptotic local power function**.

In Figure 9.2 we plot the local power function $\pi(\delta)$ as a function of $\delta \in [-1, 4]$ for tests of asymptotic size $\alpha = 0.10$, $\alpha = 0.05$, and $\alpha = 0.01$. $\delta = 0$ corresponds to the null hypothesis so $\pi(\delta) = \alpha$. The power functions are monotonically increasing in δ. Note that the power is lower than α for $\delta < 0$ due to the one-sided nature of the test.

We can see that the three power functions are ranked by α so that the test with $\alpha = 0.10$ has higher power than the test with $\alpha = 0.01$. This is the inherent trade-off between size and power. Decreasing size induces a decrease in power, and conversely.

The coefficient δ can be interpreted as the parameter deviation measured as a multiple of the standard error $s(\widehat{\theta})$. To see this, recall that $s(\widehat{\theta}) = n^{-1/2}\sqrt{\widehat{V}_\theta} \simeq n^{-1/2}\sqrt{V_\theta}$ and then note that

$$\delta = \frac{h}{\sqrt{V_\theta}} \simeq \frac{n^{-1/2}h}{s(\widehat{\theta})} = \frac{\theta_n - \theta_0}{s(\widehat{\theta})}.$$

Thus δ approximately equals the deviation $\theta_n - \theta_0$ expressed as multiples of the standard error $s(\widehat{\theta})$. Thus as we examine Figure 9.2, we can interpret the power function at $\delta = 1$ (e.g. 26% for a 5% size test) as the power when the parameter θ_n is one standard error above the hypothesized value. For example, from Table 4.1 the standard error for the coefficient on "Married Female" is 0.010. Thus in this example, $\delta = 1$ corresponds to $\theta_n = 0.010$ or an 1.0% wage premium for married females. Our calculations show that the asymptotic power of a one-sided 5% test against this alternative is about 26%.

The difference between power functions can be measured either vertically or horizontally. For example, in Figure 9.2 there is a vertical dotted line at $\delta = 1$, showing that the asymptotic local power function $\pi(\delta)$ equals 39% for $\alpha = 0.10$, equals 26% for $\alpha = 0.05$ and equals 9% for $\alpha = 0.01$. This is the difference in power across tests of differing size, holding fixed the parameter in the alternative.

A horizontal comparison can also be illuminating. To illustrate, in Figure 9.2 there is a horizontal dotted line at 50% power. 50% power is a useful benchmark, as it is the point where the test has equal odds of rejection and acceptance. The dotted line crosses the three power curves at $\delta = 1.29$ ($\alpha = 0.10$), $\delta = 1.65$ ($\alpha = 0.05$), and $\delta = 2.33$ ($\alpha = 0.01$). This means that the parameter θ must be at least 1.65 standard errors above the hypothesized value for a one-sided 5% test to have 50% (approximate) power.

The ratio of these values (e.g. $1.65/1.29 = 1.28$ for the asymptotic 5% versus 10% tests) measures the relative parameter magnitude needed to achieve the same power. (Thus, for a 5% size test to achieve 50% power, the parameter must be 28% larger than for a 10% size test.) Even more interesting, the square of this ratio (e.g. $(1.65/1.29)^2 = 1.64$) can be interpreted as the increase in sample size needed to achieve the same power under fixed parameters. That is, to achieve 50% power, a 5% size test needs 64% more observations than a 10% size test. This interpretation follows by the following informal argument. By definition and (9.25) $\delta = h/\sqrt{V_\theta} = \sqrt{n}\left(\theta_n - \theta_0\right)/\sqrt{V_\theta}$. Thus holding θ and V_θ fixed, δ^2 is proportional to n.

The analysis of a two-sided t test is similar. (9.26) implies that

$$T = \left|\frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})}\right| \xrightarrow{d} |\mathrm{Z} + \delta|$$

and thus the local power of a two-sided t test is

$$\begin{aligned}\lim_{n\to\infty} \Pr\left(\text{Reject } \mathbb{H}_0\right) &= \lim_{n\to\infty} \Pr\left(T > c\right) \\ &= \Pr\left(|\mathrm{Z} + \delta| > c\right) \\ &= \Phi\left(\delta - c\right) - \Phi\left(-\delta - c\right)\end{aligned}$$

which is monotonically increasing in $|\delta|$.

Theorem 9.22.1 *Under Assumptions 7.1.2 and 7.10.1, and* $\theta_n = r(\boldsymbol{\beta}_n) = r_0 + n^{-1/2}h$, *then*

$$T(\theta_0) = \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})} \xrightarrow{d} \mathrm{Z} + \delta$$

where $\mathrm{Z} \sim \mathrm{N}(0,1)$ *and* $\delta = h/\sqrt{V_\theta}$. *For* c *such that* $\Phi(c) = 1 - \alpha$,

$$\Pr\left(T(\theta_0) > c\right) \longrightarrow \Phi\left(\delta - c\right).$$

Furthermore, for c *such that* $\Phi(c) = 1 - \alpha/2$,

$$\Pr\left(|T(\theta_0)| > c\right) \longrightarrow \Phi\left(\delta - c\right) - \Phi\left(-\delta - c\right).$$

9.23 Asymptotic Local Power, Vector Case

In this section we extend the local power analysis of the previous section to the case of vector-valued alternatives. We generalize (9.25) to allow $\boldsymbol{\theta}_n$ to be vector-valued. The local parameterization takes the form

$$\boldsymbol{\theta}_n = \boldsymbol{r}(\boldsymbol{\beta}_n) = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h} \tag{9.27}$$

where $\boldsymbol{h}$ is $q \times 1$.

Under (9.27),

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n\right) + \boldsymbol{h} \\ &\xrightarrow{d} \mathrm{Z}_{\boldsymbol{h}} \sim \mathrm{N}(\boldsymbol{h}, \boldsymbol{V}_{\boldsymbol{\theta}}),\end{aligned}$$

a normal random vector with mean $\boldsymbol{h}$ and variance matrix $\boldsymbol{V}_{\boldsymbol{\theta}}$.

Applied to the Wald statistic we find

$$\begin{aligned}W &= n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\ &\xrightarrow{d} \mathrm{Z}_{\boldsymbol{h}}' \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \mathrm{Z}_{\boldsymbol{h}} \sim \chi_q^2(\lambda)\end{aligned} \tag{9.28}$$

where $\lambda = \boldsymbol{h}'\boldsymbol{V}^{-1}\boldsymbol{h}$. $\chi_q^2(\lambda)$ is a non-central chi-square random variable with non-centrality parameter λ. (See Section 5.3 and Theorem 5.3.3.)

The convergence (9.28) shows that under the local alternatives (9.27), $W \xrightarrow{d} \chi_q^2(\lambda)$. This generalizes the null asymptotic distribution which obtains as the special case $\lambda = 0$. We can use this result to obtain a continuous asymptotic approximation to the power function. For any significance level $\alpha > 0$ set the asymptotic critical value c so that $\Pr\left(\chi_q^2 > c\right) = \alpha$. Then as $n \to \infty$,

$$\Pr\left(W > c\right) \longrightarrow \Pr\left(\chi_q^2(\lambda) > c\right) \stackrel{def}{=} \pi(\lambda).$$

The asymptotic local power function $\pi(\lambda)$ depends only on α, q, and λ.

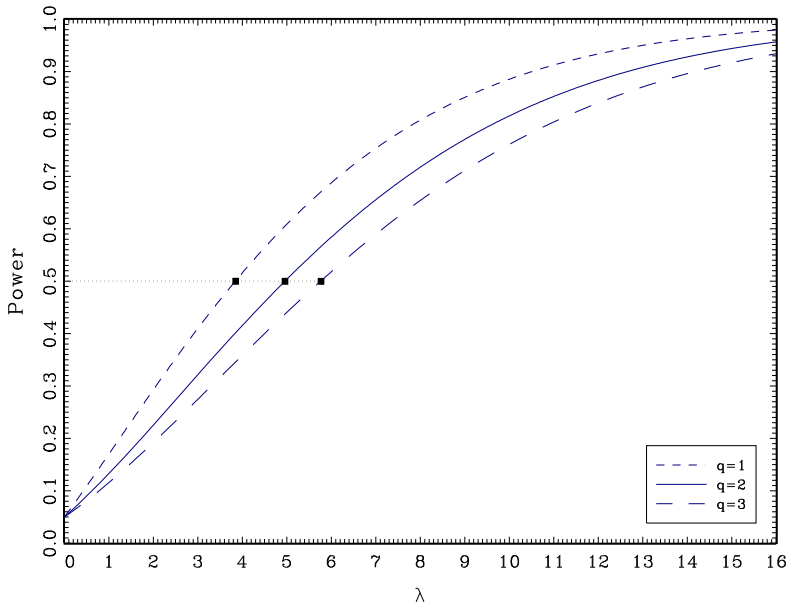


Figure 9.3: Asymptotic Local Power Function, Varying q

Theorem 9.23.1 *Under Assumptions 7.1.2 and 7.10.1, and* $\boldsymbol{\theta}_n = \boldsymbol{r}(\boldsymbol{\beta}_n) = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h}$, *then*

$$W \xrightarrow{d} \chi^2_q(\lambda)$$

where $\lambda = \boldsymbol{h}'\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{h}$. *Furthermore, for* c *such that* $\Pr\left(\chi^2_q > c\right) = \alpha$,

$$\Pr\left(W > c\right) \longrightarrow \Pr\left(\chi^2_q(\lambda) > c\right).$$

Figure 9.3 plots $\pi(\lambda)$ as a function of λ for $q = 1$, $q = 2$, and $q = 3$, and $\alpha = 0.05$. The asymptotic power functions are monotonically increasing in λ and asymptote to one.

Figure 9.3 also shows the power loss for fixed non-centrality parameter λ as the dimensionality of the test increases. The power curves shift to the right as q increases, resulting in a decrease in power. This is illustrated by the dotted line at 50% power. The dotted line crosses the three power curves at $\lambda = 3.85$ ($q = 1$), $\lambda = 4.96$ ($q = 2$), and $\lambda = 5.77$ ($q = 3$). The ratio of these λ values correspond to the relative sample sizes needed to obtain the same power. Thus increasing the dimension of the test from $q = 1$ to $q = 2$ requires a 28% increase in sample size, or an increase from $q = 1$ to $q = 3$ requires a 50% increase in sample size, to obtain a test with 50% power.

9.24 Technical Proofs*

Proof of Theorem 9.12.1. The conditions of Theorem 8.14.1 hold, since $\mathbb{H}_0$ implies Assumption 8.6.1. From (8.58) with $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$, we see that

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right) &= \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\left(\boldsymbol{R}_n^{*\prime}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}_n^{*\prime}\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &\xrightarrow{d} \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}) \\ &= \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\ \mathrm{Z}.\end{aligned}$$

where $\mathrm{Z} \sim \mathrm{N}(\mathbf{0}, (\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R})^{-1})$. Thus

$$\begin{aligned} J^* &= n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right) \\ &\xrightarrow{d} \mathrm{Z}'\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\ \mathrm{Z} \\ &= \mathrm{Z}'\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)\ \mathrm{Z} \\ &= \chi_q^2. \end{aligned}$$

■

Exercises

Exercise 9.1 Prove that if an additional regressor $\boldsymbol{X}_{k+1}$ is added to $\boldsymbol{X}$, Theil's adjusted $\overline{R}^2$ increases if and only if $|T_{k+1}| > 1$, where $T_{k+1} = \widehat{\beta}_{k+1}/s(\widehat{\beta}_{k+1})$ is the t-ratio for $\widehat{\beta}_{k+1}$ and

$$s(\widehat{\beta}_{k+1}) = \left(s^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{k+1,k+1}\right)^{1/2}$$

is the homoskedasticity-formula standard error.

Exercise 9.2 You have two independent samples $(\boldsymbol{y}_1, \boldsymbol{X}_1)$ and $(\boldsymbol{y}_2, \boldsymbol{X}_2)$ which satisfy $\boldsymbol{y}_1 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{e}_1$ and $\boldsymbol{y}_2 = \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}_2$, where $\mathbb{E}(\boldsymbol{x}_{1i}e_{1i}) = \boldsymbol{0}$ and $\mathbb{E}(\boldsymbol{x}_{2i}e_{2i}) = \boldsymbol{0}$, and both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have k columns. Let $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ be the OLS estimates of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. For simplicity, you may assume that both samples have the same number of observations n.

(a) Find the asymptotic distribution of $\sqrt{n}\left(\left(\widehat{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_1\right) - (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)\right)$ as $n \to \infty$.

(b) Find an appropriate test statistic for $\mathbb{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{\beta}_1$.

(c) Find the asymptotic distribution of this statistic under $\mathbb{H}_0$.

Exercise 9.3 Let T be a t-statistic for $\mathbb{H}_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Since $|T| \to_d |Z|$ under H_0, someone suggests the test "Reject $\mathbb{H}_0$ if $|T| < c_1$ or $|T| > c_2$, where c_1 is the $\alpha/2$ quantile of $|Z|$ and c_2 is the $1 - \alpha/2$ quantile of $|Z|$.

(a) Show that the asymptotic size of the test is α.

(b) Is this a good test of $\mathbb{H}_0$ versus $\mathbb{H}_1$? Why or why not?

Exercise 9.4 Let W be a Wald statistic for $\mathbb{H}_0 : \boldsymbol{\theta} = 0$ versus $\mathbb{H}_1 : \boldsymbol{\theta} \neq 0$, where $\boldsymbol{\theta}$ is $q \times 1$. Since $W \to_d \chi^2_q$ under H_0, someone suggests the test "Reject $\mathbb{H}_0$ if $W < c_1$ or $W > c_2$, where c_1 is the $\alpha/2$ quantile of χ^2_q and c_2 is the $1 - \alpha/2$ quantile of χ^2_q.

(a) Show that the asymptotic size of the test is α.

(b) Is this a good test of $\mathbb{H}_0$ versus $\mathbb{H}_1$? Why or why not?

Exercise 9.5 Take the linear model

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= 0 \end{aligned}$$

where both $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are $q \times 1$. Show how to test the hypotheses $\mathbb{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ against $\mathbb{H}_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$.

Exercise 9.6 Suppose a researcher wants to know which of a set of 20 regressors has an effect on a variable *testscore*. He regresses *testscore* on the 20 regressors and reports the results. One of the 20 regressors (*studytime*) has a large t-ratio (about 2.5), while other t-ratios are insignificant (smaller than 2 in absolute value). He argues that the data show that *studytime* is the key predictor for *testscore*. Do you agree with this conclusion? Is there a deficiency in his reasoning?

Exercise 9.7 Take the model

$$\begin{aligned} y_i &= x_i\beta_1 + x_i^2\beta_2 + e_i \\ \mathbb{E}(e_i \mid x_i) &= 0 \end{aligned}$$

where y_i is wages (dollars per hour) and x_i is age. Describe how you would test the hypothesis that the expected wage for a 40-year-old worker is $20 an hour.

Exercise 9.8 You want to test $\mathbb{H}_0 : \boldsymbol{\beta}_2 = 0$ against $\mathbb{H}_1 : \boldsymbol{\beta}_2 \neq 0$ in the model

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= 0 \end{aligned}$$

You read a paper which estimates model

$$y_i = \boldsymbol{x}_{1i}'\widehat{\boldsymbol{\gamma}}_1 + \left(\boldsymbol{x}_{2i} - \boldsymbol{x}_{1i}\right)'\widehat{\boldsymbol{\gamma}}_2 + \widehat{e}_i$$

and reports a test of $\mathbb{H}_0 : \boldsymbol{\gamma}_2 = 0$ against $\mathbb{H}_1 : \boldsymbol{\gamma}_2 \neq 0$. Is this related to the test you wanted to conduct?

Exercise 9.9 Suppose a researcher uses one dataset to test a specific hypothesis $\mathbb{H}_0$ against $\mathbb{H}_1$, and finds that he can reject $\mathbb{H}_0$. A second researcher gathers a similar but independent dataset, uses similar methods and finds that she cannot reject $\mathbb{H}_0$. How should we (as interested professionals) interpret these mixed results?

Exercise 9.10 In Exercise 7.8, you showed that $\sqrt{n}\left(\widehat{\sigma}^2 - \sigma^2\right) \to_d \mathrm{N}\left(0, V\right)$ as $n \to \infty$ for some V. Let $\widehat{V}$ be an estimate of V.

(a) Using this result, construct a t-statistic for $\mathbb{H}_0 : \sigma^2 = 1$ against $\mathbb{H}_1 : \sigma^2 \neq 1$.

(b) Using the Delta Method, find the asymptotic distribution of $\sqrt{n}\left(\widehat{\sigma} - \sigma\right)$.

(c) Use the previous result to construct a t-statistic for $\mathbb{H}_0 : \sigma = 1$ against $\mathbb{H}_1 : \sigma \neq 1$.

(d) Are the null hypotheses in (a) and (c) the same or are they different? Are the tests in (a) and (c) the same or are they different? If they are different, describe a context in which the two tests would give contradictory results.

Exercise 9.11 Consider a regression such as Table 4.1 where both *experience* and its square are included. A researcher wants to test the hypothesis that *experience* does not affect mean wages, and does this by computing the t-statistic for *experience*. Is this the correct approach? If not, what is the appropriate testing method?

Exercise 9.12 A researcher estimates a regression and computes a test of $\mathbb{H}_0$ against $\mathbb{H}_1$ and finds a p-value of $p = 0.08$, or "not significant". She says "I need more data. If I had a larger sample the test will have more power and then the test will reject." Is this interpretation correct?

Exercise 9.13 A common view is that "If the sample size is large enough, any hypothesis will be rejected." What does this mean? Interpret and comment.

Exercise 9.14 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= 0 \end{aligned}$$

with parameter of interest $\theta = \boldsymbol{R}'\boldsymbol{\beta}$ with $\boldsymbol{R}$ $k \times 1$. Let $\widehat{\boldsymbol{\beta}}$ be the least-squares estimate and $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ its variance estimate.

(a) Write down $\widehat{C}$, the 95% asymptotic confidence interval for θ, in terms of $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, $\boldsymbol{R}$, and $z = 1.96$ (the 97.5% quantile of $\mathrm{N}(0,1)$).

(b) Show that the decision "Reject $\mathbb{H}_0$ if $\theta_0 \notin \widehat{C}$" is an asymptotic 5% test of $\mathbb{H}_0 : \theta = \theta_0$.

Exercise 9.15 You are at a seminar where a colleague presents a simulation study of a test of a hypothesis $\mathbb{H}_0$ with nominal size 5%. Based on $B = 100$ simulation replications under $\mathbb{H}_0$ the estimated size is 7%. Your colleague says: "Unfortunately the test over-rejects."

(a) Do you agree or disagree with your colleague? Explain. Hint: Use an asymptotic (large B) approximation.

(b) Suppose the number of simulation replications were $B = 1000$ yet the estimated size is still 7%. Does your answer change?

Exercise 9.16 You have n iid observations (y_i, x_{1i}, x_{2i}), and consider two alternative regression models

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_{1i} \\ \mathbb{E}\left(\boldsymbol{x}_{1i}e_{1i}\right) &= 0 \end{aligned} \tag{9.29}$$

$$\begin{aligned} y_i &= \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_{2i} \\ \mathbb{E}\left(\boldsymbol{x}_{2i}e_{2i}\right) &= 0 \end{aligned} \tag{9.30}$$

where $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ have at least some different regressors. (For example, (9.29) is a wage regression on geographic variables and (2) is a wage regression on personal appearance measurements.) You want to know if model (9.29) or model (9.30) fits the data better. Define $\sigma_1^2 = E\left(e_{1i}^2\right)$ and $\sigma_2^2 = E\left(e_{2i}^2\right)$. You decide that the model with the smaller variance fit (e.g., model (9.29) fits better if $\sigma_1^2 < \sigma_2^2$.) You decide to test for this by testing the hypothesis of equal fit $\mathbb{H}_0 : \sigma_1^2 = \sigma_2^2$ against the alternative of unequal fit $\mathbb{H}_1 : \sigma_1^2 \neq \sigma_2^2$. For simplicity, suppose that e_{1i} and e_{2i} are observed.

(a) Construct an estimate $\widehat{\theta}$ of $\theta = \sigma_1^2 - \sigma_2^2$.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\theta} - \theta\right)$ as $n \to \infty$.

(c) Find an estimator of the asymptotic variance of $\widehat{\theta}$.

(d) Propose a test of asymptotic size α of $\mathbb{H}_0$ against $\mathbb{H}_1$.

(e) Suppose the test accepts $\mathbb{H}_0$. Briefly, what is your interpretation?

Exercise 9.17 You have two regressors x_1 and x_2, and estimate a regression with all quadratic terms

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i}^2 + \beta_4 x_{2i}^2 + \beta_5 x_{1i} x_{2i} + e_i$$

One of your advisors asks: Can we exclude the variable x_{2i} from this regression?

How do you translate this question into a statistical test? When answering these questions, be specific, not general.

(a) What is the relevant null and alternative hypotheses?

(b) What is an appropriate test statistic? Be specific.

(c) What is the appropriate asymptotic distribution for the statistic? Be specific.

(d) What is the rule for acceptance/rejection of the null hypothesis?

Exercise 9.18 The observed data is $\{y_i, \boldsymbol{x}_i, \boldsymbol{z}_i\} \in \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}^\ell$, $k > 1$ and $\ell > 1$, $i = 1, ..., n$. An econometrician first estimates

$$y_i = \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} + \widehat{e}_i$$

by least squares. The econometrician next regresses the residual $\widehat{e}_i$ on $\boldsymbol{z}_i$, which can be written as

$$\widehat{e}_i = \boldsymbol{z}_i' \widetilde{\boldsymbol{\gamma}} + \widetilde{u}_i.$$

(a) Define the population parameter $\boldsymbol{\gamma}$ being estimated in this second regression.

(b) Find the probability limit for $\widetilde{\boldsymbol{\gamma}}$.

(c) Suppose the econometrician constructs a Wald statistic W_n for $\mathbb{H}_0 : \boldsymbol{\gamma} = \mathbf{0}$ from the second regression, ignoring the regression. Write down the formula for W_n.

(d) Assuming $\mathbb{E}(\boldsymbol{z}_i \boldsymbol{x}_i') = \mathbf{0}$, find the asymptotic distribution for W_n under $\mathbb{H}_0 : \boldsymbol{\gamma} = \mathbf{0}$.

(e) If $\mathbb{E}(\boldsymbol{z}_i \boldsymbol{x}_i') \neq \mathbf{0}$ will your answer to (d) change?

Exercise 9.19 An economist estimates $y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + e_i$ by least-squares and tests the hypothesis $\mathbb{H}_0 : \beta_2 = 0$ against $\mathbb{H}_1 : \beta_2 \neq 0$. She obtains a Wald statistic $W_n = 0.34$. The sample size is $n = 500$.

(a) What is the correct degrees of freedom for the χ^2 distribution to evaluate the significance of the Wald statistic?

(b) The Wald statistic W_n is very small. Indeed, is it less than the 1% quantile of the appropriate χ^2 distribution? If so, should you reject $\mathbb{H}_0$? Explain your reasoning.

Exercise 9.20 You are reading a paper, and it reports the results from two nested OLS regressions:

$$y_i = \boldsymbol{x}_{1i}' \widetilde{\boldsymbol{\beta}}_1 + \widetilde{e}_i$$
$$y_i = \boldsymbol{x}_{1i}' \widehat{\boldsymbol{\beta}}_1 + \boldsymbol{x}_{2i}' \widehat{\boldsymbol{\beta}}_2 + \widehat{e}_i$$

Some summary statistics are reported:

Short Regression	Long Regression
$R^2 = .20$	$R^2 = .26$
$\sum_{i=1}^n \widetilde{e}_i^2 = 106$	$\sum_{i=1}^n \widehat{e}_i^2 = 100$
# of coefficients=5	# of coefficients=8
$n = 50$	$n = 50$

You are curious if the estimate $\widehat{\boldsymbol{\beta}}_2$ is statistically different from the zero vector. Is there a way to determine an answer from this information? Do you have to make any assumptions (beyond the standard regularity conditions) to justify your answer?

Exercise 9.21 Take the model

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + e_i$$
$$\mathbb{E}(\boldsymbol{x}_i e_i) = 0$$

Describe how you would test

$$\mathbb{H}_0 : \frac{\beta_1}{\beta_2} = \frac{\beta_3}{\beta_4}$$

against

$$\mathbb{H}_1 : \frac{\beta_1}{\beta_2} \neq \frac{\beta_3}{\beta_4}.$$

Exercise 9.22 You have a random sample from the model

$$\begin{aligned} y_i &= x_i\beta_1 + x_i^2\beta_2 + e_i \\ \mathbb{E}(e_i \mid x_i) &= 0 \end{aligned}$$

where y_i is wages (dollars per hour) and x_i is age. Describe how you would test the hypothesis that the expected wage for a 40-year-old worker is $20 an hour.

Exercise 9.23 Let T_n be a test statistic such that under $\mathbb{H}_0$, $T_n \to_d \chi^2_3$. Since $P\left(\chi^2_3 > 7.815\right) = 0.05$, an asymptotic 5% test of $\mathbb{H}_0$ rejects when $T_n > 7.815$. An econometrician is interested in the Type I error of this test when $n = 100$ and the data structure is well specified. She performs the following Monte Carlo experiment.

- $B = 200$ samples of size $n = 100$ are generated from a distribution satisfying $\mathbb{H}_0$.
- On each sample, the test statistic T_{nb} is calculated.
- She calculates $\widehat{p} = \frac{1}{B}\sum_{b=1}^{B} 1(T_{nb} > 7.815) = 0.070$
- The econometrician concludes that the test T_n is oversized in this context – it rejects too frequently under $\mathbb{H}_0$.

Is her conclusion correct, incorrect, or incomplete? Be specific in your answer.

Exercise 9.24 Do a Monte Carlo simulation. Take the model

$$\begin{aligned} y_i &= \alpha + x_i\beta + e_i \\ \mathbb{E}(x_i e_i) &= 0 \end{aligned}$$

where the parameter of interest is $\theta = \exp(\beta)$. Your data generating process (DGP) for the simulation is: x_i is $U[0,1]$, e_i is independent of x_i and $N(0,1)$, $n = 50$. Set $\alpha = 0$ and $\beta = 1$. Generate $B = 1000$ independent samples with α. On each, estimate the regression by least-squares, calculate the covariance matrix using a standard (heteroskedasticity-robust) formula, and similarly estimate θ and its standard error. For each replication, store $\widehat{\beta}$, $\widehat{\theta}$, $t_\beta = \left(\widehat{\beta} - \beta\right)/s\left(\widehat{\beta}\right)$, and $t_\theta = \left(\widehat{\theta} - \theta\right)/s\left(\widehat{\theta}\right)$

(a) Does the value of α matter? Explain why the described statistics are **invariant** to α and thus setting $\alpha = 0$ is irrelevant.

(b) From the 1000 replications estimate $\mathbb{E}\left(\widehat{\beta}\right)$ and $\mathbb{E}\left(\widehat{\theta}\right)$. Discuss if you see evidence if either estimator is biased or unbiased.

(c) From the 1000 replications estimate $\Pr(t_\beta > 1.645)$ and $\Pr(t_\theta > 1.645)$. What does asymptotic theory predict these probabilities should be in large samples? What do your simulation results indicate?

Exercise 9.25 The data set `invest` on the textbook website contains data on 565 U.S. firms extracted from Compustat for the year 1987. (This is one year from a panel data set used by B. E. Hansen (1999). The original data was compiled by Hall and Hall (1993).) The variables are

- I Investment to Capital Ratio (multiplied by 100).
- Q Total Market Value to Asset Ratio (Tobin's Q).

- C Cash Flow to Asset Ratio.
- D Long Term Debt to Asset Ratio.

The flow variables are annual sums for 1987. The stock variables are beginning of year.

(a) Estimate a linear regression of I_i on the other variables. Calculate appropriate standard errors.

(b) Calculate asymptotic confidence intervals for the coefficients.

(c) This regression is related to Tobin's q theory of investment, which suggests that investment should be predicted solely by Q_i. Thus the coefficient on Q_i should be positive and the others should be zero. Test the joint hypothesis that the coefficients on C_i and D_i are zero. Test the hypothesis that the coefficient on Q_i is zero. Are the results consistent with the predictions of the theory?

(d) Now try a non-linear (quadratic) specification. Regress I_i on Q_i, C_i, D_i, Q_i^2, C_i^2, D_i^2, Q_iC_i, Q_iD_i, C_iD_i. Test the joint hypothesis that the six interaction and quadratic coefficients are zero.

Exercise 9.26 In a paper in 1963, Marc Nerlove analyzed a cost function for 145 American electric companies. His data set `Nerlove1963` is on the textbook website. The variables are

- C Total cost
- Q Output
- PL Unit price of labor
- PK Unit price of capital
- PL Unit price of labor

Nerlov was interested in estimating a *cost function*: $C = f(Q, PL, PF, PK)$.

(a) First estimate an unrestricted Cobb-Douglass specification

$$\log C_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + e_i. \tag{9.31}$$

Report parameter estimates and standard errors.

(b) What is the economic meaning of the restriction $\mathbb{H}_0 : \beta_3 + \beta_4 + \beta_5 = 1$?

(c) Estimate (9.31) by constrained least-squares imposing $\beta_3+\beta_4+\beta_5 = 1$. Report your parameter estimates and standard errors.

(d) Estimate (9.31) by efficient minimum distance imposing $\beta_3 + \beta_4 + \beta_5 = 1$. Report your parameter estimates and standard errors.

(e) Test $\mathbb{H}_0 : \beta_3 + \beta_4 + \beta_5 = 1$ using a Wald statistic.

(f) Test $\mathbb{H}_0 : \beta_3 + \beta_4 + \beta_5 = 1$ using a minimum distance statistic.

Exercise 9.27 In Section 8.12 we report estimates from Mankiw, Romer and Weil (1992). We reported estimation both by unrestricted least-squares and by constrained estimation, imposing the constraint that three coefficients (2^{nd}, 3^{rd} and 4^{th} coefficients) sum to zero, as implied by the Solow growth theory. Using the same dataset `MRW1992` estimate the unrestricted model and test the hypothesis that the three coefficients sum to zero.

Exercise 9.28 Using the CPS dataset and the subsample of non-hispanic blacks (race code = 2), test the hypothesis that marriage status does not affect mean wages.

(a) Take the regression reported in Table 4.1. Which variables will need to be omitted to estimate a regression for the subsample of blacks?

(b) Express the hypothesis "marriage status does not affect mean wages" as a restriction on the coefficients. How many restrictions is this?

(c) Find the Wald (or F) statistic for this hypothesis. What is the appropriate distribution for the test statistic? Calculate the p-value of the test.

(d) What do you conclude?

Exercise 9.29 Using the CPS dataset and the subsample of non-hispanic blacks (race code = 2) and whites (race code = 1), test the hypothesis that the returns to education is common across groups.

(a) Allow the return to education to vary across the four groups (white male, white female, black male, black female) by interacting dummy variables with *education*. Estimate an appropriate version of the regression reported in Table 4.1.

(b) Find the Wald (or F) statistic for this hypothessis. What is the appropriate distribution for the test statistic? Calculate the p-value of the test.

(c) What do you conclude?

Chapter 10

Multivariate Regression

10.1 Introduction

Multivariate regression is a system of regression equations. Multivariate regression is used as reduced form models for instrumental variable estimation (explored in Chaper 11), vector autoregressions (explored in Chapter 15), demand systems (demand for multiple goods), and other contexts.

Multivariate regression is also called by the name **systems of regression equations**. Closely related is the method of **Seemingly Unrelated Regressions** (SUR) which we introduce in Section 10.7.

Most of the tools of single equation regression generalize naturally to multivariate regression. A major difference is a new set of notation to handle matrix estimates.

10.2 Regression Systems

A system of linear regressions takes the form

$$y_{ji} = \boldsymbol{x}_{ji}'\boldsymbol{\beta}_j + e_{ji} \tag{10.1}$$

for variables $j = 1, ..., m$ and observations $i = 1, ..., n$, where the regressor vectors $\boldsymbol{x}_{ji}$ are $k_j \times 1$ and e_{ji} is an error. The coefficient vectors $\boldsymbol{\beta}_j$ are $k_j \times 1$. The total number of coefficients are $\overline{k} = \sum_{j=1}^{n} k_j$. The regression system specializes to univariate regression when $m = 1$.

It is typical to treat the observations as independent across observations i but correlated across variables j. As an example, the observations y_{ji} could be expenditures by household i on good j. The standard assumptions are that households are mutually independent, but expenditures by an individual household are correlated across goods.

To describe the dependence between the dependent variables, we can define the $m \times 1$ error vector $\boldsymbol{e}_i = (e_{1i}, ..., e_{mi})'$ and its $m \times m$ variance matrix

$$\boldsymbol{\Sigma} = \mathbb{E}\left(\boldsymbol{e}_i \boldsymbol{e}_i'\right).$$

The diagonal elements are the variances of the errors e_{ji}, and the off-diagonals are the covariances across variables. It is typical to allow $\boldsymbol{\Sigma}$ to be unconstrained.

We can group the m equations (10.1) into a single equation as follows. Let $\boldsymbol{y}_i = (y_{1i}, ..., y_{mi})'$ be the $m \times 1$ vector of dependent variables, define the $\overline{k} \times m$ matrix of regressors

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_{1i} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \boldsymbol{x}_{2i} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{x}_{mi} \end{pmatrix},$$

and define the $\overline{k} \times 1$ stacked coefficient vector

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_m \end{pmatrix}.$$

Then the m regression equations can jointly be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{e}_i. \tag{10.2}$$

The entire system can be written in matrix notation by stacking the variables. Define

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix}, \qquad \boldsymbol{e} = \begin{pmatrix} \boldsymbol{e}_1 \\ \vdots \\ \boldsymbol{e}_n \end{pmatrix}, \qquad \overline{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix}$$

which are $mn \times 1$, $mn \times 1$, and $mn \times \overline{k}$, respectively. The system can be written as

$$\boldsymbol{y} = \overline{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{e}.$$

In many (perhaps most) applications the regressor vectors $\boldsymbol{x}_{ji}$ are common across the variables j, so $\boldsymbol{x}_{ji} = \boldsymbol{x}_i$ and $k_j = k$. By this we mean that the same variables enter each equation with no exclusion restrictions. Several important simplifications occur in this context. One is that we can write (10.2) using the notation

$$\boldsymbol{y}_i = \boldsymbol{B}'\boldsymbol{x}_i + \boldsymbol{e}_i$$

where $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_m)$ is $k \times m$. Another is that we can write the system in the $n \times m$ matrix notation

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1' \\ \vdots \\ \boldsymbol{y}_n' \end{pmatrix}, \qquad \boldsymbol{E} = \begin{pmatrix} \boldsymbol{e}_1' \\ \vdots \\ \boldsymbol{e}_n' \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1' \\ \vdots \\ \boldsymbol{x}_n' \end{pmatrix}.$$

Another convenient implication of common regressors is that we have the simplification

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{x}_i & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \boldsymbol{x}_i & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{x}_i \end{pmatrix} = \boldsymbol{I}_m \otimes \boldsymbol{x}_i$$

where $\otimes$ is the Kronecker product (see Appendix A.16).

10.3 Least-Squares Estimator

Consider estimating each equation (10.1) by least-squares. This takes the form

$$\widehat{\boldsymbol{\beta}}_j = \left(\sum_{i=1}^{n} \boldsymbol{x}_{ji}\boldsymbol{x}_{ji}' \right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_{ji} y_{ji} \right).$$

The combined estimate of $\boldsymbol{\beta}$ is the stacked vector

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \vdots \\ \widehat{\boldsymbol{\beta}}_m \end{pmatrix}.$$

It turns that we can write this estimator using the systems notation

$$\widehat{\boldsymbol{\beta}} = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'\boldsymbol{y}\right) = \left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{y}_i\right). \tag{10.3}$$

To see this, observe that

$$\begin{aligned}
\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}} &= \begin{pmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_n \end{pmatrix}\begin{pmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} \\
&= \sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i' \\
&= \sum_{i=1}^{n}\begin{pmatrix} \boldsymbol{x}_{1i} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \boldsymbol{x}_{2i} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{x}_{mi} \end{pmatrix}\begin{pmatrix} \boldsymbol{x}_{1i}' & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \boldsymbol{x}_{2i}' & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{x}_{mi}' \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^{n}\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}' & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \sum_{i=1}^{n}\boldsymbol{x}_{2i}\boldsymbol{x}_{2i}' & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sum_{i=1}^{n}\boldsymbol{x}_{mi}\boldsymbol{x}_{mi}' \end{pmatrix}
\end{aligned}$$

and

$$\begin{aligned}
\overline{\boldsymbol{X}}'\boldsymbol{y} &= \begin{pmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_n \end{pmatrix}\begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} \\
&= \sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{y}_i \\
&= \sum_{i=1}^{n}\begin{pmatrix} \boldsymbol{x}_{1i} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \boldsymbol{x}_{2i} & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{x}_{mi} \end{pmatrix}\begin{pmatrix} y_{1i} \\ \vdots \\ y_{mi} \end{pmatrix} \\
&= \begin{pmatrix} \sum_{i=1}^{n}\boldsymbol{x}_{1i}y_{1i} \\ \vdots \\ \sum_{i=1}^{n}\boldsymbol{x}_{mi}y_{mi} \end{pmatrix}.
\end{aligned}$$

Hence

$$\begin{aligned}
\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'\boldsymbol{y}\right) &= \left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{y}_i\right) \\
&= \begin{pmatrix} \left(\sum_{i=1}^{n}\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_{1i}y_{1i}\right) \\ \vdots \\ \left(\sum_{i=1}^{n}\boldsymbol{x}_{mi}\boldsymbol{x}_{mi}'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_{mi}y_{mi}\right) \end{pmatrix} \\
&= \widehat{\boldsymbol{\beta}}
\end{aligned}$$

as claimed.

The $m \times 1$ residual vector for the i^{th} observation is

$$\widehat{\boldsymbol{e}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$$

and the least-squares estimate of the $m \times m$ error variance matrix is

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{e}}_i \widehat{\boldsymbol{e}}_i'. \tag{10.4}$$

In the case of common regressors, observe that

$$\widehat{\boldsymbol{\beta}}_j = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{x}_i y_{ji}\right).$$

We can set

$$\widehat{\boldsymbol{B}} = \left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \cdots, \widehat{\boldsymbol{\beta}}_m\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right). \tag{10.5}$$

In Stata, multivariate regression can be implemented using the `mvreg` command.

10.4 Mean and Variance of Systems Least-Squares

We can calculate the finite-sample mean and variance of $\widehat{\boldsymbol{\beta}}$ under the conditional mean assumption

$$\mathbb{E}\left(\boldsymbol{e}_i \mid \boldsymbol{x}_i\right) = \boldsymbol{0} \tag{10.6}$$

where $\boldsymbol{x}_i$ is the union of the regressors $\boldsymbol{x}_{ji}$. Equation (10.6) is equivalent to $\mathbb{E}\left(y_{ji} \mid \boldsymbol{x}_i\right) = \boldsymbol{x}_{ji}'\boldsymbol{\beta}_j$, or that the regression model is correctly specified.

We can center the estimator as

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\overline{\boldsymbol{X}}'\boldsymbol{e}\right) = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{e}_i\right).$$

Taking conditional expectations, we find $\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$. Consequently, systems least-squares is unbiased under correct specification.

To compute the variance of the estimator, define the conditional covariance matrix of the errors of the i^{th} observation

$$\mathbb{E}\left(\boldsymbol{e}_i \boldsymbol{e}_i' \mid \boldsymbol{x}_i\right) = \boldsymbol{\Sigma}_i$$

which in general is unrestricted. Observe that if the observations are mutually independent, then

$$\begin{aligned}
\mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\begin{pmatrix} \boldsymbol{e}_1\boldsymbol{e}_1 & \boldsymbol{e}_1\boldsymbol{e}_2 & \cdots & \boldsymbol{e}_1\boldsymbol{e}_n \\ \vdots & \ddots & & \vdots \\ \boldsymbol{e}_n\boldsymbol{e}_1 & \boldsymbol{e}_n\boldsymbol{e}_2 & \cdots & \boldsymbol{e}_n\boldsymbol{e}_n \end{pmatrix} \mid \boldsymbol{X}\right) \\
&= \begin{pmatrix} \boldsymbol{\Sigma}_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\Sigma}_n \end{pmatrix}.
\end{aligned}$$

Also, by independence across observations,

$$\operatorname{var}\left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{e}_i \mid \boldsymbol{X}\right) = \sum_{i=1}^{n} \operatorname{var}\left(\boldsymbol{X}_i \boldsymbol{e}_i \mid \boldsymbol{x}_i\right) = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{\Sigma}_i \boldsymbol{X}_i'.$$

It follows that

$$\operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{\Sigma}_i \boldsymbol{X}_i'\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}.$$

When the regressors are common so that $\boldsymbol{X}_i = \boldsymbol{I}_m \otimes \boldsymbol{x}_i$ then the covariance matrix can be written as

$$\operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{I}_n \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(\sum_{i=1}^{n}\left(\boldsymbol{\Sigma}_i \otimes \boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)\left(\boldsymbol{I}_m \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

Alternatively, if the errors are conditionally homoskedastic

$$\mathbb{E}\left(\boldsymbol{e}_i\boldsymbol{e}_i' \mid \boldsymbol{x}_i\right) = \boldsymbol{\Sigma} \tag{10.7}$$

then the covariance matrix takes the form

$$\operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{\Sigma}\boldsymbol{X}_i'\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}.$$

If both simplifications (common regressors and conditional homoskedasticity) hold then we have the considerable simplication

$$\operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\Sigma} \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

10.5 Asymptotic Distribution

For an asymptotic distribution it is sufficient to consider the equation-by-equation projection model in which case

$$\mathbb{E}\left(\boldsymbol{x}_{ji}e_{ji}\right) = \boldsymbol{0}. \tag{10.8}$$

First, consider consistency. Since $\widehat{\boldsymbol{\beta}}_j$ are the standard least-squares estimators, they are consistent for the projection coefficients $\boldsymbol{\beta}_j$.

Second, consider the asymptotic distribution. Again by our single equation theory it is immediate that the $\widehat{\boldsymbol{\beta}}_j$ are asymptotically normally distributed. But our previous theory does not provide a joint distribution of the $\widehat{\boldsymbol{\beta}}_j$ across j. For this we need a joint theory for the stacked estimates $\widehat{\boldsymbol{\beta}}$, which we now provide.

Since the vector

$$\boldsymbol{X}_i\boldsymbol{e}_i = \begin{pmatrix} \boldsymbol{x}_{1i}e_{1i} \\ \vdots \\ \boldsymbol{x}_{mi}e_{mi} \end{pmatrix}$$

is i.i.d. across i and mean zero under (10.8), the central limit theorem implies

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{e}_i\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right)$$

where

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{e}_i\boldsymbol{e}_i'\boldsymbol{X}_i'\right) = \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{\Sigma}_i\boldsymbol{X}_i'\right).$$

The matrix $\boldsymbol{\Omega}$ is the covariance matrix of the variables $\boldsymbol{x}_{ji}e_{ji}$ across equations. Under conditional homoskedasticity (10.7) the matrix $\boldsymbol{\Omega}$ simplifies to

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{\Sigma}\boldsymbol{X}_i'\right) \tag{10.9}$$

(see Exercise 10.1). When the regressors are common then it simplies to

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{e}_i\boldsymbol{e}_i' \otimes \boldsymbol{x}_i\boldsymbol{x}_i'\right) \tag{10.10}$$

(see Exercise 10.2) and under both conditions (homoskedasticity and common regressors) it simplifies to

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma} \otimes \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) \tag{10.11}$$

(see Exercise 10.3).

Applied to the centered and normalized estimator we obtain the asymptotic distribution.

Theorem 10.5.1 *Under Assumption 7.1.2,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}$$

$$\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i'\right) = \begin{pmatrix} \mathbb{E}\left(\boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right) & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbb{E}\left(\boldsymbol{x}_{ni}\boldsymbol{x}_{ni}'\right) \end{pmatrix}$$

For a proof, see Exercise 10.4.

When the regressors are common then the matrix $\boldsymbol{Q}$ simplies as

$$\boldsymbol{Q} = \boldsymbol{I}_m \otimes \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) \tag{10.12}$$

(See Exercise 10.5).

If both the regressors are common and the errors are conditionally homoskedastic (10.7) then we have the simplication

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{\Sigma} \otimes \left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1} \tag{10.13}$$

(see Exercise 10.6).

Sometimes we are interested in parameters $\boldsymbol{\theta} = r(\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m) = r(\boldsymbol{\beta})$ which are functions of the coefficients from multiple equations. In this case the least-squares estimate of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}} = r(\widehat{\boldsymbol{\beta}})$. The asymptotic distribution of $\widehat{\boldsymbol{\theta}}$ can be obtained from Theorem 10.5.1 by the delta method.

Theorem 10.5.2 *Under Assumptions 7.1.2 and 7.10.1,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}$$

$$\boldsymbol{R} = \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}\left(\boldsymbol{\beta}\right)'$$

For a proof, see Exercise 10.7.

Theorem 10.5.2 is an example where multivariate regression is fundamentally distinct from univariate regression. Only by treating the least-squares estimates as a joint estimator can we obtain a distributional theory for an estimator $\widehat{\boldsymbol{\theta}}$ which is a function of estimates from multiple equations and thereby construct standard errors, confidence intervals, and hypothesis tests.

10.6 Covariance Matrix Estimation

From the finite sample and asymptotic theory we can construct appropriate estimators for the variance of $\widehat{\boldsymbol{\beta}}$. In the general case we have

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\widehat{\boldsymbol{e}}_i\widehat{\boldsymbol{e}}_i'\boldsymbol{X}_i'\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}.$$

Under conditional homoskedasticity (10.7) an appropriate estimator is

$$\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}} = \left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\widehat{\boldsymbol{\Sigma}}\boldsymbol{X}_i'\right)\left(\overline{\boldsymbol{X}}'\overline{\boldsymbol{X}}\right)^{-1}.$$

When the regressors are common then these estimators equal

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{I}_n \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(\sum_{i=1}^{n}\left(\widehat{\boldsymbol{e}}_i\widehat{\boldsymbol{e}}_i' \otimes \boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)\left(\boldsymbol{I}_n \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$$

and

$$\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}} = \widehat{\boldsymbol{\Sigma}} \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1},$$

respectively.

Covariance matrix estimators for $\widehat{\boldsymbol{\theta}}$ are found as

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}} &= \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{R}}\\ \widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\theta}}} &= \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{R}}\\ \widehat{\boldsymbol{R}} &= \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}\right)'.\end{aligned}$$

Theorem 10.6.1 *Under Assumption 7.1.2,*

$$n\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$$

and

$$n\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}} \xrightarrow{p} \boldsymbol{V}^0_{\boldsymbol{\beta}}$$

For a proof, see Exercise 10.8.

10.7 Seemingly Unrelated Regression

Consider the systems regression model under the conditional mean and conditional homoskedasticity assumptions

$$\begin{aligned}\boldsymbol{y}_i &= \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{e}_i \qquad (10.14)\\ \mathbb{E}\left(\boldsymbol{e}_i \mid \boldsymbol{x}_i\right) &= \boldsymbol{0}\\ \mathbb{E}\left(\boldsymbol{e}_i\boldsymbol{e}_i' \mid \boldsymbol{x}_i\right) &= \boldsymbol{\Sigma}\end{aligned}$$

Since the errors are correlated across equations we can consider estimation by Generalized Least Squares (GLS). To derive the estimator, premultiply (10.14) by $\boldsymbol{\Sigma}^{-1/2}$ so that the transformed error vector is i.i.d. with covariance matrix $\boldsymbol{I}_m$. Then apply least-squares and rearrange to find

$$\widehat{\boldsymbol{\beta}}_{\text{gls}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i\right). \tag{10.15}$$

(see Exercise 10.9). Another approach is to take the vector representation

$$\boldsymbol{y} = \overline{\boldsymbol{X}} \boldsymbol{\beta} + \boldsymbol{e}$$

and calculate that the equation error $\boldsymbol{e}$ has variance $\mathbb{E}(\boldsymbol{e}\boldsymbol{e}') = \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}$. Premultiply the equation by $\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}^{-1/2}$ so that the transformed error has variance matrix $\boldsymbol{I}_{nm}$ and then apply least-squares to find

$$\widehat{\boldsymbol{\beta}}_{\text{gls}} = \left(\overline{\boldsymbol{X}}' \left(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}^{-1}\right) \overline{\boldsymbol{X}}\right)^{-1} \left(\overline{\boldsymbol{X}}' \left(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{y}\right) \tag{10.16}$$

(see Exercise 10.10).

Expressions (10.15) and (10.16) are algebraically equivalent. To see the equivalence, observe that

$$\begin{aligned}
\overline{\boldsymbol{X}}' \left(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}^{-1}\right) \overline{\boldsymbol{X}} &= \begin{pmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \boldsymbol{\Sigma}^{-1} & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix} \\
&= \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_i'
\end{aligned}$$

and

$$\begin{aligned}
\overline{\boldsymbol{X}}' \left(\boldsymbol{I}_n \otimes \boldsymbol{\Sigma}^{-1}\right) \boldsymbol{y} &= \begin{pmatrix} \boldsymbol{X}_1 & \cdots & \boldsymbol{X}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \vdots & \boldsymbol{\Sigma}^{-1} & & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 \\ \vdots \\ \boldsymbol{y}_n \end{pmatrix} \\
&= \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_i.
\end{aligned}$$

Since $\boldsymbol{\Sigma}$ is unknown it must be replaced by an estimator. Using $\widehat{\boldsymbol{\Sigma}}$ from (10.4) we obtain a feasible GLS estimator.

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{sur}} &= \left(\sum_{i=1}^{n} \boldsymbol{X}_i \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{X}_i \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{y}_i\right) \\
&= \left(\overline{\boldsymbol{X}}' \left(\boldsymbol{I}_n \otimes \widehat{\boldsymbol{\Sigma}}^{-1}\right) \overline{\boldsymbol{X}}\right)^{-1} \left(\overline{\boldsymbol{X}}' \left(\boldsymbol{I}_n \otimes \widehat{\boldsymbol{\Sigma}}^{-1}\right) \boldsymbol{y}\right).
\end{aligned} \tag{10.17}$$

This is known as the **Seemingly Unrelated Regression (SUR)** estimator.

The estimator $\widehat{\boldsymbol{\Sigma}}$ can be updated by calculating the SUR residuals $\widehat{\boldsymbol{e}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i' \widehat{\boldsymbol{\beta}}_{SUR}$ and the covariance matrix estimate $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\boldsymbol{e}}_i \widehat{\boldsymbol{e}}_i'$. Substituted into (10.17) we find an iterated SUR estimator, and this can be iterated until convergence.

Under conditional homoskedasticity (10.7) we can derive its asymptotic distribution.

Theorem 10.7.1 *Under Assumption 7.1.2 and (10.7)*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{sur}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}^{*}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}}^{*} = \left(\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{\Sigma}^{-1} \boldsymbol{X}_i'\right)\right)^{-1}.$$

For a proof, see Exercise 10.11.

Under these assumptions, SUR is more efficient than least-squares (in particular, under the assumption of conditional homoskedasticity).

Theorem 10.7.2 *Under Assumption 7.1.2 and (10.7)*

$$\begin{aligned} \boldsymbol{V}_{\boldsymbol{\beta}}^{*} &= \left(\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i'\right)\right)^{-1} \\ &\leq \left(\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{\Sigma}\boldsymbol{X}_i'\right)\left(\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i'\right)\right)^{-1} \\ &= \boldsymbol{V}_{\boldsymbol{\beta}} \end{aligned}$$

and thus $\widehat{\boldsymbol{\beta}}_{\text{sur}}$ *is asymptotically more efficient than* $\widehat{\boldsymbol{\beta}}_{OLS}$.

For a proof, see Exercise 10.12.

An appropriate estimator of the variance of $\widehat{\boldsymbol{\beta}}_{SUR}$ is

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \left(\sum_{i=1}^{n}\boldsymbol{X}_i\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_i'\right)^{-1}.$$

Theorem 10.7.3 *Under Assumption 7.1.2 and (10.7)*

$$n\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$$

and thus $\widehat{\boldsymbol{\beta}}_{SUR}$ *is asymptotically more efficient than* $\widehat{\boldsymbol{\beta}}_{OLS}$.

For a proof, see Exercise 10.13.

In Stata, the seemingly unrelated regressions estimator is implemented using the `sureg` command.

Arnold Zellner

Arnold Zellner (1927-2000) of the United States was a founding father of the econometrics field. He was a pioneer in Bayesian econometrics. One of his core contributions was the method of Seemingly Unrelated Regressions.

10.8 Maximum Likelihood Estimator

Take the linear model under the assumption that the error is independent of the regressors and multivariate normally distributed. Thus

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{X}_i'\boldsymbol{\beta} + \boldsymbol{e}_i \\ \boldsymbol{e}_i &\sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right). \end{aligned}$$

In this case we can consider the maximum likelihood estimator (MLE) of the coefficients.

It is convenient to reparameterize the covariance matrix in terms of its inverse, thus $\boldsymbol{S} = \boldsymbol{\Sigma}^{-1}$. With this reparameterization, the conditional denstiy of $\boldsymbol{y}_i$ given $\boldsymbol{X}_i$ equals

$$f\left(\boldsymbol{y}_i|\boldsymbol{X}_i\right) = \frac{\det\left(\boldsymbol{S}\right)^{1/2}}{(2\pi)^{m/2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta}\right)' \boldsymbol{S}\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta}\right)\right).$$

The log-likelihood function for the sample is

$$\log L(\boldsymbol{\beta}, \boldsymbol{S}) = -\frac{nm}{2}\log\left(2\pi\right) + \frac{n}{2}\log\det\left(\boldsymbol{S}\right) - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta}\right)' \boldsymbol{S}\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta}\right).$$

The maximum likelihood estimator $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{S}}\right)$ maximizes the log-likelihood function. The first order conditions are

$$\begin{aligned} 0 &= \frac{\partial}{\partial \boldsymbol{\beta}} \log L(\boldsymbol{\beta}, \boldsymbol{S})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{S}=\widehat{\boldsymbol{S}}} \\ &= \sum_{i=1}^{n} \boldsymbol{X}_i \widehat{\boldsymbol{S}}\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}\right) \end{aligned}$$

and

$$\begin{aligned} 0 &= \frac{\partial}{\partial \boldsymbol{S}} \log L(\boldsymbol{\beta}, \boldsymbol{\Sigma})\Big|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}, \boldsymbol{S}=\widehat{\boldsymbol{S}}} \\ &= \frac{n}{2}\widehat{\boldsymbol{S}}^{-1} - \frac{1}{2}\operatorname{tr}\left(\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}\right)'\right). \end{aligned}$$

The second equation uses the matrix results $\frac{\partial}{\partial \boldsymbol{S}}\log\det\left(\boldsymbol{S}\right) = \boldsymbol{S}^{-1}$ and $\frac{\partial}{\partial \boldsymbol{B}}\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{B}\right) = \boldsymbol{A}'$ from Appendix A.15.

Solving and making the substitution $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{S}}^{-1}$ we obtain

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{X}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{X}_i \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{y}_i\right)$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}\right)'.$$

Notice that each equation refers to the other. Hence these are not closed-form expressions, but can be solved via iteration. The solution is identical to the iterated SUR estimator. Thus the SUR estimator (iterated) is identical to the MLE under normality.

Recall that the SUR estimator simplifies to OLS when the regressors are common across equations. The same occurs for the MLE. Thus when $\boldsymbol{X}_i = \boldsymbol{I}_m \otimes \boldsymbol{x}_i$ we find that $\widehat{\boldsymbol{\beta}}_{MLE} = \widehat{\boldsymbol{\beta}}_{OLS}$ and $\widehat{\boldsymbol{\Sigma}}_{MLE} = \widehat{\boldsymbol{\Sigma}}_{OLS}$.

10.9 Reduced Rank Regression

One context where systems estimation is important is when it is desired to impose or test restrictions across equations. Restricted systems are commonly estimated by maximum likelihood under normality. In this section we explore one important special case of restricted multivariate regression known as reduced rank regression. The model was originally proposed by Anderson (1951) and extended by Johansen (1995).

The unrestricted model is

$$\begin{aligned} \boldsymbol{y}_i &= \boldsymbol{B}'\boldsymbol{x}_i + \boldsymbol{C}'\boldsymbol{z}_i + \boldsymbol{e}_i \\ \mathbb{E}\left(\boldsymbol{e}_i\boldsymbol{e}_i' \mid \boldsymbol{x}_i, \boldsymbol{z}_i\right) &= \boldsymbol{\Sigma} \end{aligned} \tag{10.18}$$

where $\boldsymbol{B}$ is $k \times m$, $\boldsymbol{C}$ is $\ell \times m$, and $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are regressors. We separate the regressors $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ because the coefficient matrix $\boldsymbol{B}$ will be restricted while $\boldsymbol{C}$ will be unrestricted.

The matrix $\boldsymbol{B}$ is full rank if

$$\operatorname{rank}(\boldsymbol{B}) = \min(k, m).$$

The reduced rank restriction is that

$$\operatorname{rank}(\boldsymbol{B}) = r < \min(k, m)$$

for some known r.

The reduced rank restriction implies that we can write the coefficient matrix $\boldsymbol{B}$ in the factored form

$$\boldsymbol{B} = \boldsymbol{G}\boldsymbol{A}' \tag{10.19}$$

where $\boldsymbol{A}$ is $m \times r$ and $\boldsymbol{G}$ is $k \times r$. This representation is not unique (as we can replace $\boldsymbol{G}$ with $\boldsymbol{G}\boldsymbol{Q}$ and $\boldsymbol{A}$ with $\boldsymbol{A}\boldsymbol{Q}^{-1\prime}$ for any invertible $\boldsymbol{Q}$ and the same relation holds). Identification therefore requires a normalization of the coefficients. A conventional normalization is

$$\boldsymbol{G}'\boldsymbol{D}\boldsymbol{G} = \boldsymbol{I}_r$$

for given $\boldsymbol{D}$.

Equivalently, the reduced rank restriction can be imposed by requiring that $\boldsymbol{B}$ satisfy the restriction $\boldsymbol{B}\boldsymbol{A}_\perp = \boldsymbol{G}\boldsymbol{A}'\boldsymbol{A}_\perp = \boldsymbol{0}$ for some $m \times (m-r)$ coefficient matrix $\boldsymbol{A}_\perp$. Since $\boldsymbol{G}$ is full rank this requires that $\boldsymbol{A}'\boldsymbol{A}_\perp = \boldsymbol{0}$, hence $\boldsymbol{A}_\perp$ is the orthogonal complement to $\boldsymbol{A}$. Note that $\boldsymbol{A}_\perp$ is not unique as it can be replaced by $\boldsymbol{A}_\perp\boldsymbol{Q}$ for any $(m-r) \times (m-r)$ invertible $\boldsymbol{Q}$. Thus if $\boldsymbol{A}_\perp$ is to be estimated it requires a normalization.

We discuss methods for estimation of $\boldsymbol{G}$, $\boldsymbol{A}$, $\boldsymbol{\Sigma}$, $\boldsymbol{C}$, and $\boldsymbol{A}_\perp$. The standard approach is maximum likelihood under the assumption that $\boldsymbol{e}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. The log-likelihood function for the sample is

$$\begin{aligned} \log L(\boldsymbol{G}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{\Sigma}) = &-\frac{nm}{2}\log(2\pi) - \frac{n}{2}\log\det(\boldsymbol{\Sigma}) \\ &- \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \boldsymbol{A}\boldsymbol{G}'\boldsymbol{x}_i - \boldsymbol{C}'\boldsymbol{z}_i\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{A}\boldsymbol{G}'\boldsymbol{x}_i - \boldsymbol{C}'\boldsymbol{z}_i\right). \end{aligned}$$

Anderson (1951) derived the MLE by imposing the constraint $\boldsymbol{B}\boldsymbol{A}_\perp = \boldsymbol{0}$ via the method of Lagrange multipliers. This turns out to be algebraically cumbersome.

Johansen (1995) instead proposed a concentration method which turns out to be relatively straightforward. The method is as follows. First, treat $\boldsymbol{G}$ as if it is known. Then maximize the log-likelihood with respect to the other parameters. Resubstituting these estimates, we obtain the concentrated log-likelihood function with respect to $\boldsymbol{G}$. This can be maximized to find the MLE for $\boldsymbol{G}$. The other parameter estimates are then obtain by substitution. We now describe these steps in detail.

Given $\boldsymbol{G}$, the likelihood is a normal multivariate regression in the variables $\boldsymbol{G}'\boldsymbol{x}_i$ and $\boldsymbol{z}_i$, so the MLE for $\boldsymbol{A}$, $\boldsymbol{C}$ and $\boldsymbol{\Sigma}$ are least-squares. In particular, using the Frisch-Waugh-Lovell residual regression formula, we can write the estimators for $\boldsymbol{A}$ and $\boldsymbol{\Sigma}$ as

$$\widehat{\boldsymbol{A}}(\boldsymbol{G}) = \left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)^{-1}$$

and

$$\widehat{\boldsymbol{\Sigma}}(\boldsymbol{G}) = \frac{1}{n}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\right)$$

where

$$\widetilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}$$
$$\widetilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}.$$

Substituting these estimators into the log-likelihood function, we obtain the concentrated likelihood function, which is a function of $\boldsymbol{G}$ only

$$\begin{aligned}\log \widetilde{L}(\boldsymbol{G}) &= \log L\left(\boldsymbol{G}, \widehat{\boldsymbol{A}}(\boldsymbol{G}), \widehat{\boldsymbol{C}}(\boldsymbol{G}), \widehat{\boldsymbol{\Sigma}}(\boldsymbol{G})\right) \\ &= \frac{m}{2}\left(n\log(2\pi) - 1\right) - \frac{n}{2}\log\det\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\right) \\ &= \frac{m}{2}\left(n\log(2\pi) - 1\right) - \frac{n}{2}\log\det\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)\frac{\det\left(\boldsymbol{G}'\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}\right)\boldsymbol{G}\right)}{\det\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)}.\end{aligned}$$

The third equality uses Theorem A.7.1.8. The MLE $\widehat{\boldsymbol{G}}$ for $\boldsymbol{G}$ is the maximizer of $\log\widetilde{L}(\boldsymbol{G})$, or equivalently equals

$$\widehat{\boldsymbol{G}} = \underset{\boldsymbol{G}}{\operatorname{argmin}}\frac{\det\left(\boldsymbol{G}'\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}\right)\boldsymbol{G}\right)}{\det\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)} \tag{10.20}$$

$$= \underset{\boldsymbol{G}}{\operatorname{argmax}}\frac{\det\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)}{\det\left(\boldsymbol{G}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\boldsymbol{G}\right)} \tag{10.21}$$

$$= \{\boldsymbol{v}_1, ..., \boldsymbol{v}_r\}$$

which are the generalized eigenvectors of $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}$ with respect to $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}$ corresponding to the r largest generalized eigenvalues. (Generalized eigenvalues and eigenvectors are discussed in Section A.10.) The estimator satisfies the normalization $\widehat{\boldsymbol{G}}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{G}} = \boldsymbol{I}_r$. Letting $\boldsymbol{v}_j^*$ denote the eigenvectors of (10.20), we can also express $\widehat{\boldsymbol{G}} = \left\{\boldsymbol{v}_m^*, ..., \boldsymbol{v}_{m-r+1}^*\right\}$.

This is computationally straightforward. In MATLAB, for example, the generalized eigenvalues and eigenvectors of a matrix $\boldsymbol{A}$ with respect to $\boldsymbol{B}$ are found using the command `eig(A,B)`.

Given $\widehat{\boldsymbol{G}}$, the MLE $\widehat{\boldsymbol{A}}$, $\widehat{\boldsymbol{C}}$, $\widehat{\boldsymbol{\Sigma}}$ are found by least-squares regression of $\boldsymbol{y}_i$ on $\widehat{\boldsymbol{G}}'\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. In particular, $\widehat{\boldsymbol{A}} = \widehat{\boldsymbol{G}}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}$ since $\widehat{\boldsymbol{G}}'\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{G}} = \boldsymbol{I}_r$.

We now discuss the estimator $\widehat{\boldsymbol{A}}_\perp$ of $\boldsymbol{A}_\perp$. It turns out that

$$\widehat{\boldsymbol{A}}_\perp = \underset{\boldsymbol{A}}{\operatorname{argmax}}\frac{\det\left(\boldsymbol{A}'\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\right)\boldsymbol{A}\right)}{\det\left(\boldsymbol{A}'\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\boldsymbol{A}\right)} \tag{10.22}$$

$$= \{\boldsymbol{w}_1, ..., \boldsymbol{w}_{m-r}\}$$

the eigenvectors of $\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}$ with respect to $\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}$ associated with the largest $m - r$ eigenvalues.

By the dual eigenvalue relation (Theorem A.10.1), the eigenvalue problems in equations (10.20) and (10.22) have the same non-unit eigenvalues λ_j, and the associated eigenvectors $\boldsymbol{v}_j^*$ and $\boldsymbol{w}_j$ satisfy

the relationship $\boldsymbol{w}_j = \lambda_j^{-1/2}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\boldsymbol{v}_j^*$. Letting $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_m, ..., \lambda_{m-r+1}\}$ this implies

$$\begin{aligned}\{\boldsymbol{w}_m, ..., \boldsymbol{w}_{m-r+1}\} &= \left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\left\{\boldsymbol{v}_m^*, ..., \boldsymbol{v}_{m-r+1}^*\right\}\boldsymbol{\Lambda} \\ &= \left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\widehat{\boldsymbol{A}}\boldsymbol{\Lambda}.\end{aligned}$$

The second equality holds since $\widehat{\boldsymbol{G}} = \left\{\boldsymbol{v}_m^*, ..., \boldsymbol{v}_{m-r+1}^*\right\}$ and $\widehat{\boldsymbol{A}} = \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\widehat{\boldsymbol{G}}$. Since the eigenvectors $\boldsymbol{w}_j$ satisfy the orthogonality property $\boldsymbol{w}_j'\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\boldsymbol{w}_\ell = 0$ for $j \neq \ell$, it follows that

$$0 = \widehat{\boldsymbol{A}}_\perp'\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\{\boldsymbol{w}_m, ..., \boldsymbol{w}_{m-r+1}\} = \widehat{\boldsymbol{A}}_\perp'\widehat{\boldsymbol{A}}\boldsymbol{\Lambda}.$$

Since $\boldsymbol{\Lambda} > 0$ we conclude that $\widehat{\boldsymbol{A}}_\perp'\widehat{\boldsymbol{A}} = 0$ as desired.

The solution $\widehat{\boldsymbol{A}}_\perp$ in (10.22) can be represented several ways. One which is computationally convenient is to observe that

$$\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}}\left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{X}} = \boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{Y} = \widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}}$$

where $\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}} = \boldsymbol{I}_n - (\boldsymbol{X}, \boldsymbol{Z})\left((\boldsymbol{X}, \boldsymbol{Z})'(\boldsymbol{X}, \boldsymbol{Z})\right)^{-1}(\boldsymbol{X}, \boldsymbol{Z})'$ and $\widetilde{\boldsymbol{e}} = \boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{Y}$ is the residual from the unrestricted least-squares regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$. The first equality follows by the Frisch-Waugh-Lovell theorem. This shows that $\widehat{\boldsymbol{A}}_\perp$ are the generalized eigenvectors of $\widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}}$ with respect to $\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}$ corresponding to the $m - r$ largest eigenvalues. In MATLAB, for example, these can be computed using the `eig(A,B)` command.

Another representation is to write $\boldsymbol{M}_{\boldsymbol{Z}} = \boldsymbol{I}_n - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$ so that

$$\widehat{\boldsymbol{A}}_\perp = \underset{\boldsymbol{A}}{\mathrm{argmax}}\frac{\det\left(\boldsymbol{A}'\boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{Y}\boldsymbol{A}\right)}{\det\left(\boldsymbol{A}'\boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{Y}\boldsymbol{A}\right)} = \underset{\boldsymbol{A}}{\mathrm{argmin}}\frac{\det\left(\boldsymbol{A}'\boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{Y}\boldsymbol{A}\right)}{\det\left(\boldsymbol{A}'\boldsymbol{Y}'\boldsymbol{M}_{\boldsymbol{X},\boldsymbol{Z}}\boldsymbol{Y}\boldsymbol{A}\right)}$$

We summarize our findings.

Theorem 10.9.1 *The MLE for the reduced rank model (10.18) under $\boldsymbol{e}_i \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ is given as follows. $\widehat{\boldsymbol{G}} = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_r\}$, the generalized eigenvectors of $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}}\left(\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}\right)^{-1}\boldsymbol{Y}'\widetilde{\boldsymbol{X}}$ with respect to $\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}$ corresponding to the r largest eigenvalues. $\widehat{\boldsymbol{A}}$, $\widehat{\boldsymbol{C}}$ and $\widehat{\boldsymbol{\Sigma}}$ are obtained by the least-squares regression*

$$\begin{aligned}\boldsymbol{y}_i &= \widehat{\boldsymbol{A}}\widehat{\boldsymbol{G}}'\boldsymbol{x}_i + \widehat{\boldsymbol{C}}'\boldsymbol{z}_i + \widehat{\boldsymbol{e}}_i \\ \widehat{\boldsymbol{\Sigma}} &= \frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{e}}_i\widehat{\boldsymbol{e}}_i'.\end{aligned}$$

$\widehat{\boldsymbol{A}}_\perp$ equals the generalized eigenvectors of $\widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}}$ with respect to $\widetilde{\boldsymbol{Y}}'\widetilde{\boldsymbol{Y}}$ corresponding to the $m - r$ smallest eigenvalues.

Exercises

Exercise 10.1 Show (10.9) when the errors are conditionally homoskedastic (10.7).

Exercise 10.2 Show (10.10) when the regressors are common across equations $\boldsymbol{x}_{ji} = \boldsymbol{x}_i$

Exercise 10.3 Show (10.11) when the regressors are common across equations $\boldsymbol{x}_{ji} = \boldsymbol{x}_i$ and the errors are conditionally homoskedastic (10.7).

Exercise 10.4 Prove Theorem 10.5.1.

Exercise 10.5 Show (10.12) when the regressors are common across equations $\boldsymbol{x}_{ji} = \boldsymbol{x}_i$

Exercise 10.6 Show (10.13) when the regressors are common across equations $\boldsymbol{x}_{ji} = \boldsymbol{x}_i$ and the errors are conditionally homoskedastic (10.7).

Exercise 10.7 Prove Theorem 10.5.2.

Exercise 10.8 Prove Theorem 10.6.1.

Exercise 10.9 Show that (10.15) follows from the steps described.

Exercise 10.10 Show that (10.16) follows from the steps described.

Exercise 10.11 Prove Theorem 10.7.1.

Exercise 10.12 Prove Theorem 10.7.2.

Hint: First, show that it is sufficient to show that

$$\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i'\right)\left(\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{X}_i'\right) \leq \mathbb{E}\left(\boldsymbol{X}_i\boldsymbol{\Sigma}\boldsymbol{X}_i'\right).$$

Second, rewrite this equation using the transformations $\boldsymbol{U}_i = \boldsymbol{X}_i\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{V}_i = \boldsymbol{X}_i\boldsymbol{\Sigma}^{-1/2}$, and then apply the matrix Cauchy-Schwarz inequality (B.11).

Exercise 10.13 Prove Theorem 10.7.3

Exercise 10.14 Take the model

$$\begin{aligned} y_i &= \boldsymbol{\pi}_i'\boldsymbol{\beta} + e_i \\ \boldsymbol{\pi}_i &= \mathbb{E}\left(\boldsymbol{x}_i|\boldsymbol{z}_i\right) = \boldsymbol{\Gamma}'\boldsymbol{z}_i \\ \mathbb{E}\left(e_i|z_i\right) &= 0 \end{aligned}$$

where y_i, *is* scalar, $\boldsymbol{x}_i$ is a k vector and $\boldsymbol{z}_i$ is an ℓ vector. $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_i$ are $k \times 1$ and $\boldsymbol{\Gamma}$ is $\ell \times k$. The sample is $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i : i = 1, ..., n)$ with $\boldsymbol{\pi}_i$ unobserved.

Consider the estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ by OLS of y_i on $\widehat{\boldsymbol{\pi}}_i = \widehat{\boldsymbol{\Gamma}}'\boldsymbol{z}_i$ where $\widehat{\boldsymbol{\Gamma}}$ is the OLS coefficient from the multivariate regression of $\boldsymbol{x}_i$ on $\boldsymbol{z}_i$

(a) Show that $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$

(b) Find the asymptotic distribution $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ as $n \to \infty$ assuming that $\boldsymbol{\beta} = \mathbf{0}$.

(c) Why is the assumption $\boldsymbol{\beta} = \mathbf{0}$ an important simplifying condition in part (b)?

(d) Using the result in (c), construct an appropriate asymptotic test for the hypothesis $\mathbb{H}_0 : \boldsymbol{\beta} = \mathbf{0}$.

Exercise 10.15 The observations are iid, $(y_{1i}, y_{2i}, \boldsymbol{x}_i : i = 1, ..., n)$. The dependent variables y_{1i} and y_{2i} are real-valued. The regressor $\boldsymbol{x}_i$ is a k-vector. The model is the two-equation system

$$\begin{aligned} y_{1i} &= \boldsymbol{x}_i'\boldsymbol{\beta}_1 + e_{1i} \\ \mathbb{E}\left(\boldsymbol{x}_i e_{1i}\right) &= 0 \\ y_{2i} &= x_i'\boldsymbol{\beta}_2 + e_{2i} \\ \mathbb{E}\left(\boldsymbol{x}_i e_{2i}\right) &= 0 \end{aligned}$$

(a) What are the appropriate estimators $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$?

(b) Find the joint asymptotic distribution of $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$.

(c) Describe a test for $\mathbb{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

Chapter 11

Instrumental Variables

11.1 Introduction

We say that there is **endogeneity** in the linear model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{11.1}$$

if $\boldsymbol{\beta}$ is the parameter of interest and

$$\mathbb{E}(\boldsymbol{x}_i e_i) \neq \mathbf{0}. \tag{11.2}$$

This is a core problem in econometrics and largely differentiates econometrics from many branches of statistics. To distinguish (11.1) from the regression and projection models, we will call (11.1) a **structural equation** and $\boldsymbol{\beta}$ a **structural parameter**. When (11.2) holds, it is typical to say that $\boldsymbol{x}_i$ is **endogenous** for $\boldsymbol{\beta}$.

Endogeneity cannot happen if the coefficient is defined by linear projection. Indeed, we can define the linear projection coefficient $\boldsymbol{\beta}^* = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i y_i\right)$ and linear projection equation

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta}^* + e_i^* \\ \mathbb{E}(\boldsymbol{x}_i e_i^*) &= \mathbf{0}. \end{aligned}$$

However, under endogeneity (11.2) the projection coefficient $\boldsymbol{\beta}^*$ does not equal the structural parameter. Indeed,

$$\begin{aligned} \boldsymbol{\beta}^* &= \left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i y_i\right) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i\left(\boldsymbol{x}_i'\boldsymbol{\beta} + e_i\right)\right) \\ &= \boldsymbol{\beta} + \left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i e_i\right) \\ &\neq \boldsymbol{\beta} \end{aligned}$$

the final relation since $\mathbb{E}\left(\boldsymbol{x}_i e_i\right) \neq \mathbf{0}$.

Thus endogeneity requires that the coefficient be defined differently than projection. We describe such definitions as **structural**. We will present three examples in the following section.

Endogeneity implies that the least-squares estimator is inconsistent for the structural parameter. Indeed, under i.i.d. sampling, least-squares is consistent for the projection coefficient, and thus is inconsistent for $\boldsymbol{\beta}$.

$$\widehat{\boldsymbol{\beta}} \xrightarrow{p} \left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i y_i\right) = \boldsymbol{\beta}^* \neq \boldsymbol{\beta}.$$

The inconsistency of least-squares is typically referred to as **endogeneity bias** or **estimation bias** due to endogeneity. (This is an imperfect label as the actual issue is inconsistency, not bias.)

As the structural parameter $\boldsymbol{\beta}$ is the parameter of interest, endogeneity requires the development of alternative estimation methods. We discuss those in later sections.

11.2 Examples

The concept of endogeneity may be easiest to understand by example. We discuss three distinct examples. In each case it is important to see how the structural parameter $\boldsymbol{\beta}$ is defined independently from the linear projection model.

Example: Measurement error in the regressor. Suppose that $(y_i, \boldsymbol{z}_i)$ are joint random variables, $\mathbb{E}(y_i \mid \boldsymbol{z}_i) = \boldsymbol{z}_i'\boldsymbol{\beta}$ is linear, $\boldsymbol{\beta}$ is the structural parameter, and $\boldsymbol{z}_i$ is not observed. Instead we observe $\boldsymbol{x}_i = \boldsymbol{z}_i + \boldsymbol{u}_i$ where $\boldsymbol{u}_i$ is a $k \times 1$ measurement error, independent of e_i and $\boldsymbol{z}_i$. This is an example of a latent variable model, where "latent" refers to a structural variable which is unobserved.

The model $\boldsymbol{x}_i = \boldsymbol{z}_i + \boldsymbol{u}_i$ with $\boldsymbol{z}_i$ and $\boldsymbol{u}_i$ independent and $\mathbb{E}(\boldsymbol{u}_i) = \mathbf{0}$ is known as **classical measurement error**. This means that $\boldsymbol{x}_i$ is a noisy but unbiased measure of $\boldsymbol{z}_i$.

By substitution we can express y_i as a function of the observed variable $\boldsymbol{x}_i$.

$$\begin{aligned} y_i &= \boldsymbol{z}_i'\boldsymbol{\beta} + e_i \\ &= (\boldsymbol{x}_i - \boldsymbol{u}_i)'\boldsymbol{\beta} + e_i \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + v_i \end{aligned}$$

where $v_i = e_i - \boldsymbol{u}_i'\boldsymbol{\beta}$. This means that $(y_i, \boldsymbol{x}_i)$ satisfy the linear equation

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + v_i$$

with an error v_i. But this error is not a projection error. Indeed,

$$\mathbb{E}(\boldsymbol{x}_i v_i) = \mathbb{E}\left[(\boldsymbol{z}_i + \boldsymbol{u}_i)\left(e_i - \boldsymbol{u}_i'\boldsymbol{\beta}\right)\right] = -\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)\boldsymbol{\beta} \neq \mathbf{0}$$

if $\boldsymbol{\beta} \neq 0$ and $\mathbb{E}(\boldsymbol{u}_i\boldsymbol{u}_i') \neq 0$. As we learned in the previous section, if $\mathbb{E}(\boldsymbol{x}_i v_i) \neq 0$ then least-squares estimation will be inconsistent.

We can calculate the form of the projection coefficient (which is consistently estimated by least-squares). For simplicity suppose that $k = 1$. We find

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} + \frac{\mathbb{E}(x_i v_i)}{\mathbb{E}\left(x_i^2\right)} = \boldsymbol{\beta}\left(1 - \frac{\mathbb{E}\left(u_i^2\right)}{\mathbb{E}\left(x_i^2\right)}\right).$$

Since $\mathbb{E}\left(u_i^2\right)/\mathbb{E}\left(x_i^2\right) < 1$ the projection coefficient shrinks the structural parameter $\boldsymbol{\beta}$ towards zero. This is called **measurement error bias** or **attenuation bias**.

Example: Supply and Demand. The variables q_i and p_i (quantity and price) are determined jointly by the demand equation

$$q_i = -\beta_1 p_i + e_{1i}$$

and the supply equation

$$q_i = \beta_2 p_i + e_{2i}.$$

Assume that $\boldsymbol{e}_i = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$ is i.i.d., $\mathbb{E}(\boldsymbol{e}_i) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{e}_i\boldsymbol{e}_i') = \boldsymbol{I}_2$ (the latter for simplicity). The question is: if we regress q_i on p_i, what happens?

It is helpful to solve for q_i and p_i in terms of the errors. In matrix notation,

$$\begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}\begin{pmatrix} q_i \\ p_i \end{pmatrix} = \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix}$$

so

$$\begin{aligned}\begin{pmatrix} q_i \\ p_i \end{pmatrix} &= \begin{bmatrix} 1 & \beta_1 \\ 1 & -\beta_2 \end{bmatrix}^{-1} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \\ &= \begin{bmatrix} \beta_2 & \beta_1 \\ 1 & -1 \end{bmatrix} \begin{pmatrix} e_{1i} \\ e_{2i} \end{pmatrix} \left(\frac{1}{\beta_1+\beta_2}\right) \\ &= \begin{pmatrix} (\beta_2 e_{1i} + \beta_1 e_{2i})/(\beta_1+\beta_2) \\ (e_{1i} - e_{2i})/(\beta_1+\beta_2) \end{pmatrix}.\end{aligned}$$

The projection of q_i on p_i yields

$$\begin{aligned} q_i &= \beta^* p_i + e_i^* \\ \mathbb{E}(p_i e_i^*) &= 0 \end{aligned}$$

where

$$\beta^* = \frac{\mathbb{E}(p_i q_i)}{\mathbb{E}(p_i^2)} = \frac{\beta_2 - \beta_1}{2}.$$

Thus the projection coefficient β^* equals neither the demand slope β_1 nor the supply slope β_2, but equals an average of the two. (The fact that it is a simple average is an artifact of the simple covariance structure.)

Hence the OLS estimate satisfies $\widehat{\beta} \xrightarrow{p} \beta^*$, and the limit does not equal either β_1 or β_2. The fact that the limit is neither the supply nor demand slope is called **simultaneous equations bias**. This occurs generally when y_i and x_i are jointly determined, as in a market equilibrium.

Generally, when both the dependent variable and a regressor are simultaneously determined, then the variables should be treated as endogenous.

Example: Choice Variables as Regressors. Take the classic wage equation

$$\log(wage) = \beta education + e$$

with β the average causal effect of education on wages. If wages are affected by unobserved ability, and individuals with high ability self-select into higher education, then e contains unobserved ability, so *education* and e will be positively correlated. Hence *education* is endogenous. The positive correlation means that the linear projection coefficient β^* will be upward biased relative to the structural coefficient β. Thus least-squares (which is estimating the projection coefficient) will tend to over-estimate the causal effect of education on wages.

This type of endogeneity occurs generally when y and x are both choices made by an economic agent, even if they are made at different points in time.

Generally, when both the dependent variable and a regressor are choice variables made by the same agent, the variables should be treated as endogenous.

11.3 Instrumental Variables

We have defined endogeneity as the context where the regressor is correlated with the equation error. In most applications we only treat a subset of the regressors as endogenous; most of the regressors will be treated as **exogenous**, meaning that they are assumed uncorrelated with the equation error. To be specific, we make the partition

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{x}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix} \tag{11.3}$$

and similarly

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}$$

so that the **structural equation** is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i. \end{aligned} \tag{11.4}$$

The regressors are assumed to satisfy

$$\begin{aligned} \mathbb{E}(\boldsymbol{x}_{1i}e_i) &= \mathbf{0} \\ \mathbb{E}(\boldsymbol{x}_{2i}e_i) &\neq \mathbf{0}. \end{aligned}$$

We call $\boldsymbol{x}_{1i}$ **exogenous** and $\boldsymbol{x}_{2i}$ **endogenous** for the structural parameter $\boldsymbol{\beta}$. As the dependent variable y_i is also endogenous, we sometimes differentiate $\boldsymbol{x}_{2i}$ by calling $\boldsymbol{x}_{2i}$ the **endogenous right-hand-side variables**.

In matrix notation we can write (11.4) as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}. \end{aligned} \tag{11.5}$$

The endogenous regressors $\boldsymbol{x}_{2i}$ are the critical variables discussed in the examples of the previous section – simultaneous variables, choice variables, mis-measured regressors – that are potentially correlated with the equation error e_i. In most applications the number k_2 of variables treated as endogenous is small (1 or 2). The exogenous variables $\boldsymbol{x}_{1i}$ are the remaining regressors (including the equation intercept) and can be low or high dimensional.

To consistently estimate $\boldsymbol{\beta}$ we require additional information. One type of information which is commonly used in economic applications are what we call **instruments**.

Definition 11.3.1 *The $\ell \times 1$ random vector $\boldsymbol{z}_i$ is an **instrumental variable** for (11.4) if*

$$\mathbb{E}\left(\boldsymbol{z}_i e_i\right) = \mathbf{0} \tag{11.6}$$

$$\mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i'\right) > 0 \tag{11.7}$$

$$\operatorname{rank}\left(\mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right)\right) = k. \tag{11.8}$$

There are three components to the definition as given. The first (11.6) is that the instruments are uncorrelated with the regression error. The second (11.7) is a normalization which excludes linearly redundant instruments. The third (11.8) is often called the **relevance condition** and is essential for the identification of the model, as we discuss later. A necessary condition for (11.8) is that $\ell \geq k$.

Condition (11.6) – that the instruments are uncorrelated with the equation error, is often described as that they are **exogenous** in the sense that they are determined outside the model for y_i.

Notice that the regressors $\boldsymbol{x}_{1i}$ satisfy condition (11.6) and thus should be included as instrumental variables. It is thus a subset of the variables $\boldsymbol{z}_i$. Notationally we make the partition

$$\boldsymbol{z}_i = \begin{pmatrix} \boldsymbol{z}_{1i} \\ \boldsymbol{z}_{2i} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{1i} \\ \boldsymbol{z}_{2i} \end{pmatrix} \begin{matrix} k_1 \\ \ell_2 \end{matrix}. \tag{11.9}$$

Here, $\boldsymbol{x}_{1i} = \boldsymbol{z}_{1i}$ are the **included exogenous variables**, and $\boldsymbol{z}_{2i}$ are the **excluded exogenous variables**. That is, $\boldsymbol{z}_{2i}$ are variables which could be included in the equation for y_i (in the sense

that they are uncorrelated with e_i) yet can be excluded, as they would have true zero coefficients in the equation.

Many authors simply label $\boldsymbol{x}_{1i}$ as the "exogenous variables", $\boldsymbol{x}_{2i}$ as the "endogenous variables", and $\boldsymbol{z}_{2i}$ as the "instrumental variables".

We say that the model is **just-identified** if $\ell = k$ (and $\ell_2 = k_2$) and **over-identified** if $\ell > k$ (and $\ell_2 > k_2$).

What variables can be used as instrumental variables? From the definition $\mathbb{E}(\boldsymbol{z}_i e_i) = \mathbf{0}$ we see that the instrument must be uncorrelated with the equation error, meaning that it is excluded from the structural equation as mentioned above. From the rank condition (11.8) it is also important that the instrumental variable be correlated with the endogenous variables $\boldsymbol{x}_{2i}$ after controlling for the other exogenous variables $\boldsymbol{x}_{1i}$. These two requirements are typically interpreted as requiring that the instruments be determined outside the system for $(y_i, \boldsymbol{x}_{2i})$, causally determine $\boldsymbol{x}_{2i}$, but do not causally determine y_i except through $\boldsymbol{x}_{2i}$.

Let's take the three examples given above.

Measurement error in the regressor. When $\boldsymbol{x}_i$ is a mis-measured version of $\boldsymbol{z}_i$, a common choice for an instrument $\boldsymbol{z}_{2i}$ is an alternative measurement of $\boldsymbol{z}_i$. For this $\boldsymbol{z}_{2i}$ to satisfy the property of an instrumental variable the measurement error in $\boldsymbol{z}_{2i}$ must be independent of that in $\boldsymbol{x}_i$.

Supply and Demand. An appropriate instrument for price p_i in a demand equation is a variable z_{2i} which influences supply but not demand. Such a variable affects the equilibrium values of p_i and q_i but does not directly affect price except through quantity. Variables which affect supply but not demand are typically related to production costs.

An appropriate instrument for price in a supply equation is a variable which influences demand but not supply. Such a variable affects the equilibrium values of price and quantity but only affects price through quantity.

Choice Variable as Regressor. An ideal instrument affects the choice of the regressor (education) but does not directly influence the dependent variable (wages) except through the indirect effect on the regressor. We will discuss an example in the next section.

11.4 Example: College Proximity

In a influential paper, David Card (1995) suggested if a potential student lives close to a college, this reduces the cost of attendence and thereby raises the likelihood that the student will attend college. However, college proximity does not directly affect a student's skills or abilities, so should not have a direct effect on his or her market wage. These considerations suggest that college proximity can be used as an instrument for education in a wage regression. We use the simplist model reported in Card's paper to illustrate the concepts of instrumental variables throughout the chapter.

Card used data from the National Longitudinal Survey of Young Men (NLSYM) for 1976. A baseline least-squares wage regression for his data set is reported in the first column of Table 11.1. The dependent variable is the log of weekly earnings. The regressors are *education* (years of schooling), *experience* (years of work experience, calculated as *age* (years) less *education+6*), *experience*^{2}/100, *black*, *south* (an indicator for residence in the southern region of the U.S.), and *urban* (an indicator for residence in a standard metropolitan statistical area). We drop observations for which *wage* is missing. The remaining sample has 3,010 observations. His data is the file `Card1995` on the textbook website.

The point estimate obtained by least-squares suggests an 8% increase in earnings for each year of education.

Table 11.1
Dependent variable *log(wage)*

	OLS	IV(a)	IV(b)	2SLS(a)	2SLS(b)	LIML
education	0.074	0.132	0.133	0.161	0.160	0.164
	(0.004)	(0.049)	(0.051)	(0.040)	(0.041)	(0.042)
experience	0.084	0.107	0.056	0.119	0.047	0.120
	(0.007)	(0.021)	(0.026)	(0.018)	(0.025)	(0.019)
experience^2/100	−0.224	−0.228	−0.080	−0.231	−0.032	−0.231
	(0.032)	(0.035)	(0.133)	(0.037)	(0.127)	(0.037)
black	−0.190	−0.131	−0.103	−0.102	−0.064	−0.099
	(0.017)	(0.051)	(0.075)	(0.044)	(0.061)	(0.045)
south	−0.125	−0.105	−0.098	−0.095	−0.086	−0.094
	(0.015)	(0.023)	(0.0287)	(0.022)	(0.026)	(0.022)
urban	0.161	0.131	0.108	0.116	0.083	0.115
	(0.015)	(0.030)	(0.049)	(0.026)	(0.041)	(0.027)
Sargan				0.82	0.52	0.82
p-value				0.36	0.47	0.37

Notes:

1. IV(a) uses *college* as an instrument for *education*.
2. IV(b) uses *college*, *age*, and age^2 as instruments for *education*, *experience*, and $experience^2/100$.
3. 2SLS(a) uses *public* and *private* as instruments for *education*.
4. 2SLS(b) uses *public*, *private*, *age*, and age^2 as instruments for *education*, *experience*, and $experience^2/100$.
5. LIML uses *public* and *private* as instruments for *education*.

As discussed in the previous sections, it is reasonable to view years of education as a choice made by an individual, and thus is likely endogenous for the structural return to education. This means that least-squares is an estimate of a linear projection, but is inconsistent for coefficient of a structural equation representing the causal impact of years of education on expected wages. Labor economics predicts that ability, education, and wages will be positively correlated. This suggests that the population projection coefficient estimated by leat-squares will be higher than the structural parameter (and hence upwards biased). However, the sign of the bias is uncertain since there are multiple regressors and there are other potential sources of endogeneity.

To instrument for the endogeneity of education, Card suggested that a reasonable instrument is a dummy variable indicating if the individual grew up near a college. We will consider three measures:

college	Grew up in same county as a 4-year college
public	Grew up in same county as a 4-year public college
private	Grew up in same county as a 4-year private college.

David Card

David Card (1956-) is a Canadian-American labor economist whose research has changed our understanding of labor markets, the impact of minimum wage legislation, and immigration. His methodological innovations in applied econometrics have transformed empirical microeconomics.

11.5 Reduced Form

The reduced form is the relationship between the regressors $\boldsymbol{x}_i$ and the instruments $\boldsymbol{z}_i$. A linear reduced form model for $\boldsymbol{x}_i$ is

$$\boldsymbol{x}_i = \boldsymbol{\Gamma}'\boldsymbol{z}_i + \boldsymbol{u}_i. \tag{11.10}$$

This is a multivariate regression as introduced in Chapter 10. The $\ell \times k$ coefficient matrix $\boldsymbol{\Gamma}$ can be defined by linear projection. Thus

$$\boldsymbol{\Gamma} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i'\right) \tag{11.11}$$

so that

$$\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{u}_i'\right) = \boldsymbol{0}.$$

In matrix notation, we can write (11.10) as

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U} \tag{11.12}$$

where $\boldsymbol{U}$ is $n \times k$. Notice that the projection coefficient (11.11) is well defined and unique under (11.7).

Since $\boldsymbol{z}_i$ and $\boldsymbol{x}_i$ have the common variables $\boldsymbol{x}_{1i}$, we can focus on the reduced form for the the endogenous regressors $\boldsymbol{x}_{2i}$. Recalling the partitions (11.3) and (11.9) we can partition $\boldsymbol{\Gamma}$ conformably as

$$\begin{aligned}\boldsymbol{\Gamma} &= \begin{array}{c} \\ \\ \end{array}\begin{array}{c} \begin{array}{cc} k_1 & k_2 \end{array} \\ \begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{bmatrix} \end{array}\begin{array}{c} \\ \ell_1 \\ \ell_2 \end{array} \\ &= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{0} & \boldsymbol{\Gamma}_{22} \end{bmatrix}\end{aligned} \tag{11.13}$$

and similarly partition $\boldsymbol{u}_i$. Then (11.10) can be rewritten as two equation systems

$$\boldsymbol{x}_{1i} = \boldsymbol{z}_{1i} \tag{11.14}$$

$$\boldsymbol{x}_{2i} = \boldsymbol{\Gamma}_{12}'\boldsymbol{z}_{1i} + \boldsymbol{\Gamma}_{22}'\boldsymbol{z}_{2i} + \boldsymbol{u}_{2i}. \tag{11.15}$$

The first equation (11.14) is a tautology. The second equation (11.15) is the primary reduced form equation of interest. It is a multivariate linear regression for $\boldsymbol{x}_{2i}$ as a function of the included and excluded exogeneous variables $\boldsymbol{z}_{1i}$ and $\boldsymbol{z}_{2i}$.

We can also construct a reduced form equation for y_i. Substituting (11.10) into (11.4), we find

$$\begin{aligned} y_i &= \left(\boldsymbol{\Gamma}'\boldsymbol{z}_i + \boldsymbol{u}_i\right)'\boldsymbol{\beta} + e_i \\ &= \boldsymbol{z}_i'\boldsymbol{\lambda} + v_i \end{aligned} \tag{11.16}$$

where

$$\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta} \tag{11.17}$$

and

$$v_i = \boldsymbol{u}_i'\boldsymbol{\beta} + e_i.$$

Observe that

$$\mathbb{E}\left(\boldsymbol{z}_i v_i\right) = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{u}_i'\right)\boldsymbol{\beta} + \mathbb{E}\left(\boldsymbol{z}_i e_i\right) = \boldsymbol{0}.$$

Thus (11.16) is a projection equation. It is the reduced form for y_i, as it expresses y_i as a function of exogeneous variables only. Since it is a projection equation we can write the reduced form coefficient as

$$\boldsymbol{\lambda} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_i y_i\right) \tag{11.18}$$

which is well defined under (11.7).

Alternatively, we can substitute (11.15) into (11.4) and use $\boldsymbol{x}_{1i} = \boldsymbol{z}_{1i}$ to obtain

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \left(\boldsymbol{\Gamma}_{12}'\boldsymbol{z}_{1i} + \boldsymbol{\Gamma}_{22}'\boldsymbol{z}_{2i} + \boldsymbol{u}_{2i}\right)'\boldsymbol{\beta}_2 + e_i \\ &= \boldsymbol{z}_{1i}'\boldsymbol{\lambda}_1 + \boldsymbol{z}_{2i}'\boldsymbol{\lambda}_2 + v_i \end{aligned} \tag{11.19}$$

where

$$\boldsymbol{\lambda}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}_{12}\boldsymbol{\beta}_2 \tag{11.20}$$

$$\boldsymbol{\lambda}_2 = \boldsymbol{\Gamma}_{22}\boldsymbol{\beta}_2. \tag{11.21}$$

which is an alternative (and equivalent) expression of (11.17) given (11.13).

(11.10) and (11.16) together (or (11.15) and (11.19) together) are the **reduced form equations** for the system

$$\begin{aligned} y_i &= \boldsymbol{z}_i'\boldsymbol{\lambda} + v_i \\ \boldsymbol{x}_i &= \boldsymbol{\Gamma}'\boldsymbol{z}_i + \boldsymbol{u}_i. \end{aligned}$$

The relationships (11.17) and (11.20)-(11.21) are critically important for understanding the identification of the structural parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, as we discuss below. These equations show the tight relationship between the parameters of the structural equations ($\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$) and those of the reduced form equations ($\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, $\boldsymbol{\Gamma}_{12}$ and $\boldsymbol{\Gamma}_{22}$).

11.6 Reduced Form Estimation

The reduced form equations are projections, so the coefficient matrices may be estimated by least-squares (see Chapter 10). The least-squares estimate of (11.10) is

$$\widehat{\boldsymbol{\Gamma}} = \left(\sum_{i=1}^{n} \boldsymbol{z}_i\boldsymbol{z}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{z}_i\boldsymbol{x}_i'\right). \tag{11.22}$$

The estimates of equation (11.10) can be written as

$$\boldsymbol{x}_i = \widehat{\boldsymbol{\Gamma}}'\boldsymbol{z}_i + \widehat{\boldsymbol{u}}_i. \tag{11.23}$$

In matrix notation, these can be written as

$$\widehat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$$

and

$$\boldsymbol{X} = \boldsymbol{Z}\widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{U}}.$$

Since $\boldsymbol{X}$ and $\boldsymbol{Z}$ have a common sub-matrix, we have the partition

$$\widehat{\boldsymbol{\Gamma}} = \begin{bmatrix} \boldsymbol{I} & \widehat{\boldsymbol{\Gamma}}_{12} \\ \boldsymbol{0} & \widehat{\boldsymbol{\Gamma}}_{22} \end{bmatrix}.$$

The reduced form estimates of equation (11.15) can be written as

$$\boldsymbol{x}_{2i} = \widehat{\boldsymbol{\Gamma}}_{12}'\boldsymbol{z}_{1i} + \widehat{\boldsymbol{\Gamma}}_{22}'\boldsymbol{z}_{2i} + \widehat{\boldsymbol{u}}_{2i}$$

or in matrix notation as

$$\boldsymbol{X}_2 = \boldsymbol{Z}_1\widehat{\boldsymbol{\Gamma}}_{12} + \boldsymbol{Z}_2\widehat{\boldsymbol{\Gamma}}_{22} + \widehat{\boldsymbol{U}}_2.$$

We can write the submatrix estimates as

$$\begin{bmatrix} \widehat{\boldsymbol{\Gamma}}_{12} \\ \widehat{\boldsymbol{\Gamma}}_{22} \end{bmatrix} = \left(\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{x}_{2i}'\right) = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \left(\boldsymbol{Z}'\boldsymbol{X}_2\right).$$

The reduced form estimate of equation (11.16) is

$$\begin{aligned} \widehat{\boldsymbol{\lambda}} &= \left(\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{z}_i y_i\right) \\ y_i &= \boldsymbol{z}_i' \widehat{\boldsymbol{\lambda}} + \widehat{v}_i \\ &= \boldsymbol{z}_{1i}' \widehat{\boldsymbol{\lambda}}_1 + \boldsymbol{z}_{2i}' \widehat{\boldsymbol{\lambda}}_2 + \widehat{v}_i \end{aligned}$$

or in matrix notation

$$\begin{aligned} \widehat{\boldsymbol{\lambda}} &= \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \left(\boldsymbol{Z}'\boldsymbol{y}\right) \\ \boldsymbol{y} &= \boldsymbol{Z}\widehat{\boldsymbol{\lambda}} + \widehat{\boldsymbol{v}} \\ &= \boldsymbol{Z}_1 \widehat{\boldsymbol{\lambda}}_1 + \boldsymbol{Z}_2 \widehat{\boldsymbol{\lambda}}_2 + \widehat{\boldsymbol{v}}. \end{aligned}$$

11.7 Identification

A parameter is **identified** if it is a unique function of the probability distribution of the observables. One way to show that a parameter is identified is to write it as an explicit function of population moments. For example, the reduced form coefficient matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\lambda}$ are identified since they can be written as explicit functions of the moments of the observables $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. That is,

$$\boldsymbol{\Gamma} = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i'\right)^{-1} \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right) \tag{11.24}$$

$$\boldsymbol{\lambda} = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i'\right)^{-1} \mathbb{E}\left(\boldsymbol{z}_i y_i\right). \tag{11.25}$$

These are uniquely determined by the probability distribution of $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ if Definition 11.3.1 holds, since this includes the requirement that $\mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i'\right)$ is invertible.

We are interested in the structural parameter $\boldsymbol{\beta}$. It relates to $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$ through (11.17), or

$$\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta}. \tag{11.26}$$

It is identified if it uniquely determined by this relation. This is a set of ℓ equations with k unknowns with $\ell \geq k$. From standard linear algebra we know that there is a unique solution if and only if $\boldsymbol{\Gamma}$ has full rank k.

$$\operatorname{rank}\left(\boldsymbol{\Gamma}\right) = k. \tag{11.27}$$

Under (11.27), $\boldsymbol{\beta}$ can be uniquely solved from the linear system $\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta}$. On the other hand if $\operatorname{rank}\left(\boldsymbol{\Gamma}\right) < k$ then $\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta}$ has fewer mutually independent linear equations than coefficients so there is not a unique solution.

From the definitions (11.24)-(11.25) the identification equation (11.26) is the same as

$$\mathbb{E}\left(\boldsymbol{z}_i y_i\right) = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right) \boldsymbol{\beta}$$

which is again a set of ℓ equations with k unknowns. This has a unique solution if (and only if)

$$\operatorname{rank}\left(\mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right)\right) = k \tag{11.28}$$

which was listed in (11.8) as a conditions of Definition 11.3.1. (Indeed, this is why it was listed as part of the definition.) We can also see that (11.27) and (11.28) are equivalent ways of expressing the

same requirement. If this condition fails then $\boldsymbol{\beta}$ will not be identified. The condition (11.27)-(11.28) is called the **relevance condition**.

It is useful to have explicit expressions for the solution $\boldsymbol{\beta}$. The easiest case is when $\ell = k$. Then (11.27) implies $\boldsymbol{\Gamma}$ is invertible, so the structural parameter equals $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}$. It is a unique solution because $\boldsymbol{\Gamma}$ and $\boldsymbol{\lambda}$ are unique and $\boldsymbol{\Gamma}$ is invertible.

When $\ell > k$ we can solve for $\boldsymbol{\beta}$ by applying least-squares to the system of equations $\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta}$. This is ℓ equations with k unknowns and no error. The least-squares solution is $\boldsymbol{\beta} = \left(\boldsymbol{\Gamma}'\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\boldsymbol{\lambda}$. Under (11.27) the matrix $\boldsymbol{\Gamma}'\boldsymbol{\Gamma}$ is invertible so the solution is unique.

$\boldsymbol{\beta}$ is identified if $\operatorname{rank}(\boldsymbol{\Gamma}) = k$, which is true if and only if $\operatorname{rank}(\boldsymbol{\Gamma}_{22}) = k_2$ (by the upper-diagonal structure of $\boldsymbol{\Gamma}$). Thus the key to identification of the model rests on the $\ell_2 \times k_2$ matrix $\boldsymbol{\Gamma}_{22}$ in (11.15). To see this, recall the reduced form relationships (11.20)-(11.21). We can see that $\boldsymbol{\beta}_2$ is identified from (11.21) alone, and the necessary and sufficient condition is $\operatorname{rank}(\boldsymbol{\Gamma}_{22}) = k_2$. If this is satisfied then the solution can be written as $\boldsymbol{\beta}_2 = \left(\boldsymbol{\Gamma}_{22}'\boldsymbol{\Gamma}_{22}\right)^{-1}\boldsymbol{\Gamma}_{22}'\boldsymbol{\lambda}_2$. Then $\boldsymbol{\beta}_1$ is identified from this and (11.20), with the explicit solution $\boldsymbol{\beta}_1 = \boldsymbol{\lambda}_1 - \boldsymbol{\Gamma}_{12}\left(\boldsymbol{\Gamma}_{22}'\boldsymbol{\Gamma}_{22}\right)^{-1}\boldsymbol{\Gamma}_{22}'\boldsymbol{\lambda}_2$. In the just-identified case ($\ell_2 = k_2$) these equations simplify to take the form $\boldsymbol{\beta}_2 = \boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{\lambda}_2$ and $\boldsymbol{\beta}_1 = \boldsymbol{\lambda}_1 - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{\lambda}_2$.

11.8 Instrumental Variables Estimator

In this section we consider the special case where the model is just-identified, so that $\ell = k$.

The assumption that $\boldsymbol{z}_i$ is an instrumental variable implies that

$$\mathbb{E}\left(\boldsymbol{z}_i e_i\right) = \mathbf{0}.$$

Making the substitution $e_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$ we find

$$\mathbb{E}\left(\boldsymbol{z}_i\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right) = \mathbf{0}.$$

Expanding,

$$\mathbb{E}\left(\boldsymbol{z}_i y_i\right) - \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i'\right)\boldsymbol{\beta} = \mathbf{0}.$$

This is a system of $\ell = k$ equations and k unknowns. Solving for $\boldsymbol{\beta}$ we find

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_i y_i\right).$$

This solution assumes that the matrix $\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i'\right)$ is invertible, which holds under (11.8) or equivalently (11.27).

The **instrumental variables** (IV) estimator $\boldsymbol{\beta}$ replaces the population moments by their sample versions. We find

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{iv}} &= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i y_i\right) \\
&= \left(\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{z}_i y_i\right) \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right). \qquad (11.29)
\end{aligned}$$

More generally, it is common to refer to any estimator of the form

$$\widehat{\boldsymbol{\beta}}_{\text{iv}} = \left(\boldsymbol{W}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{W}'\boldsymbol{y}\right)$$

given an $n \times k$ matrix $\boldsymbol{W}$ as an IV estimator for $\boldsymbol{\beta}$ using the instrument $\boldsymbol{W}$.

Alternatively, recall that when $\ell = k$ the structural parameter can be written as a function of the reduced form parameters as $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}$. Replacing $\boldsymbol{\Gamma}$ and $\boldsymbol{\lambda}$ by their least-squares estimates we can construct what is called the **Indirect Least Squares** (ILS) estimator:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{\text{ils}} &= \widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\lambda}} \\
&= \left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right)\right) \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right) \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right).
\end{aligned}$$

We see that this equals the IV estimator (11.29). Thus the ILS and IV estimators are equivalent.

Given the IV estimator we define the residual vector

$$\widehat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{iv}}$$

which satisfies

$$\boldsymbol{Z}'\widehat{\boldsymbol{e}} = \boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{X}\left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right) = \boldsymbol{0}. \tag{11.30}$$

Since $\boldsymbol{Z}$ includes an intercept, this means that the residuals sum to zero, and are uncorrelated with the included and excluded instruments.

To illustrate, we estimate the reduced form equations corresponding to the college proximity example of Table 11.1, now treating *education* as endogenous and using *college* as an instrumental variable. The reduced form equations for *log(wage)* and *education* are reported in the first and second columns of Table 11.2.

Table 11.2
Reduced Form Regressions

	log(wage)	education	education	experience	experience^2/100	education
experience	0.053	−0.410				−0.413
	(0.007)	(0.032)				(0.032)
experience^2/100	−0.219	0.073				0.093
	(0.033)	(0.170)				(0.171)
black	−0.264	−1.006	−1.468	1.468	0.282	−1.006
	(0.018)	(0.088)	(0.115)	(0.115)	(0.026)	(0.088)
south	−0.143	−0.291	−0.460	0.460	0.112	−0.267
	(0.017)	(0.078)	(0.103)	(0.103)	(0.022)	(0.079)
urban	0.185	0.404	0.835	−0.835	−0.176	0.400
	(0.017)	(0.085)	(0.112)	(0.112)	(0.025)	(0.085)
college	0.045	0.337	0.347	−0.347	−0.073	
	(0.016)	(0.081)	(0.109)	(0.109)	(0.023)	
public						0.430
						(0.086)
private						0.123
						(0.101)
age			1.061	−0.061	−0.555	
			(0.296)	(0.296)	(0.065)	
age^2/100			−1.876	1.876	1.313	
			(0.516)	(0.516)	(0.116)	
F		17.51	8.22	1581	1112	13.87

Of particular interest is the equation for the endogenous regressor (education), and the coefficients for the excluded instruments – in this case *college*. The estimated coefficient equals 0.346

with a small standard error. This implies that growing up near a 4-year college increases average educational attainment by 0.3 years. This seems to be a reasonable magnitude.

Since the structural equation is just-identified with one right-hand-side endogenous variable, we can calculate the ILS/IV estimate for the education coefficient as the ratio of the coefficient estimates for the instrument *college* in the two equations, e.g. $0.346/0.047 = 0.135$, implying a 13% return to each year of education. This is substantially greater than the 8% least-squares estimate from the first column of Table 11.1.

The IV estimates of the full equation are reported in the second column of Table 11.1.

Card (1995) also points out that if *education* is endogenous, then so is our measure of *experience*, since it is calculated by subtracting *education* from *age*. He suggests that we can use the variables *age* and age^2 as instruments for *experience* and $experience^2$, as they are clearly exogeneous and yet highly correlated with *experience* and $experience^2$. Notice that this approach treats $experience^2$ as a variable separate from *experience*. Indeed, this is the correct approach.

Following this recommendation we now have three endogenous regressors and three instruments. We present the three reduced form equations for the three endogenous regressors in the third through fifth columns of Table 11.2. It is interesting to compare the equations for *education* and *experience*. The two sets of coefficients are simply the sign change of the other, with the exception of the coefficient on *age*. Indeed this must be the case, because the three variables are linearly related. Does this cause a problem for 2SLS? Fortunately, no. The fact that the coefficient on age is not simply a sign change means that the equations are not linearly singular. Hence Assumption (11.27) is not violated.

The IV estimates using the three instruments *college*, *age* and age^2 for the endogenous regressors *education*, *experience* and $experience^2$ is presented in the third column of Table 11.1. The estimate of the returns to schooling is not affected by this change in the instrument set, but the estimated return to experience profile flattens (the quadratic effect diminishes).

The IV estimator may be calculated in Stata using the `ivregress 2sls` command.

11.9 Demeaned Representation

Does the well-known demeaned representation for linear regression (3.20) carry over to the IV estimator? To see this, write the linear projection equation in the format

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \alpha + e_i$$

where α is the intercept and $\boldsymbol{x}_i$ does not contain a constant. Similarly, partition the instrument as $(1, \boldsymbol{z}_i)$ where $\boldsymbol{z}_i$ does not contain an intercept. We can write the IV estimates as

$$y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{iv}} + \widehat{\alpha}_{\text{iv}} + \widehat{e}_i$$

The orthogonality (11.30) implies the two-equation system

$$\sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{iv}} - \widehat{\alpha}_{\text{iv}}\right) = 0$$

$$\sum_{i=1}^{n}\boldsymbol{z}_i\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{iv}} - \widehat{\alpha}_{\text{iv}}\right) = \mathbf{0}.$$

The first equation implies

$$\widehat{\alpha}_{\text{iv}} = \overline{y} - \overline{\boldsymbol{x}}'\widehat{\boldsymbol{\beta}}_{\text{iv}}.$$

Substituting into the second equation

$$\sum_{i=1}^{n}\boldsymbol{z}_i\left((y_i - \overline{y}) - (\boldsymbol{x}_i - \overline{\boldsymbol{x}})'\,\widehat{\boldsymbol{\beta}}_{\text{iv}}\right)$$

and solving for $\widehat{\boldsymbol{\beta}}_{\text{iv}}$ we find

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{iv}} &= \left(\sum_{i=1}^{n} \boldsymbol{z}_i \left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)'\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{z}_i \left(y_i - \overline{y}\right)\right) \\ &= \left(\sum_{i=1}^{n} \left(\boldsymbol{z}_i - \overline{\boldsymbol{z}}\right) \left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)'\right)^{-1} \left(\sum_{i=1}^{n} \left(\boldsymbol{z}_i - \overline{\boldsymbol{z}}\right) \left(y_i - \overline{y}\right)\right). \end{aligned} \tag{11.31}$$

Thus the demeaning equations for least-squares carry over to the IV estimator. The coefficient estimate $\widehat{\boldsymbol{\beta}}_{\text{iv}}$ is a function only of the demeaned data.

11.10 Wald Estimator

In many cases, including the Card proximity example, the excluded instrument is a binary (dummy) variable. Let's focus on that case, and suppose that the model has just one endogenous regressor and no other regressors beyond the intercept. Thus the model can be written as

$$\begin{aligned} y_i &= x_i\beta + \alpha + e_i \\ \mathbb{E}\left(e_i \mid z_i\right) &= 0 \end{aligned}$$

with z_i binary.

Notice that if we take expectations of the structural equation given $z_i = 1$ and $z_i = 0$, respectively, we obtain

$$\begin{aligned} \mathbb{E}\left(y_i \mid z_i = 1\right) &= \mathbb{E}\left(x_i \mid z_i = 1\right)\beta + \alpha \\ \mathbb{E}\left(y_i \mid z_i = 0\right) &= \mathbb{E}\left(x_i \mid z_i = 0\right)\beta + \alpha. \end{aligned}$$

Subtracting and dividing, we obtain an expression for the slope coefficient β

$$\beta = \frac{\mathbb{E}\left(y_i \mid z_i = 1\right) - \mathbb{E}\left(y_i \mid z_i = 0\right)}{\mathbb{E}\left(x_i \mid z_i = 1\right) - \mathbb{E}\left(x_i \mid z_i = 0\right)}. \tag{11.32}$$

The natural moment estimator for β replaces the expectations by the averages within the "grouped data" where $z_i = 1$ and $z_i = 0$, respectively. That is, define the group means

$$\begin{aligned} \overline{y}_1 &= \frac{\sum_{i=1}^{n} z_i y_i}{\sum_{i=1}^{n} z_i}, \qquad \overline{y}_0 = \frac{\sum_{i=1}^{n} \left(1 - z_i\right) y_i}{\sum_{i=1}^{n} \left(1 - z_i\right)} \\ \overline{x}_1 &= \frac{\sum_{i=1}^{n} z_i x_i}{\sum_{i=1}^{n} z_i}, \qquad \overline{x}_0 = \frac{\sum_{i=1}^{n} \left(1 - z_i\right) x_i}{\sum_{i=1}^{n} \left(1 - z_i\right)} \end{aligned}$$

and the moment estimator

$$\widehat{\beta} = \frac{\overline{y}_1 - \overline{y}_0}{\overline{x}_1 - \overline{x}_0}. \tag{11.33}$$

This is known as the "Wald estimator" as it was proposed by Wald (1940).

These expressions are rather insightful. (11.32) shows that the structural slope coefficient is the expected change in y_i due to changing the instrument divided by the expected change in x_i due to changing the instrument. Informally, it is the change in y (due to z) over the change in x (due to z). Equation (11.33) shows that slope coefficient can be estimated by a simple ratio in means.

The expression (11.33) may appear like a distinct estimator from the IV estimator $\widehat{\boldsymbol{\beta}}_{\text{iv}}$, but it turns out that they are the same. That is, $\widehat{\beta} = \widehat{\boldsymbol{\beta}}_{\text{iv}}$. To see this, use (11.31) to find

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{iv}} &= \frac{\sum_{i=1}^{n} z_i \left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} z_i \left(x_i - \overline{x}\right)} \\ &= \frac{\overline{y}_1 - \overline{y}}{\overline{x}_1 - \overline{x}}. \end{aligned}$$

Then notice

$$\overline{y}_1 - \overline{y} = \overline{y}_1 - \left(\frac{1}{n}\sum_{i=1}^{n} z_i \overline{y}_1 + \frac{1}{n}\sum_{i=1}^{n} (1 - z_i)\, \overline{y}_0 \right) = \frac{1}{n}\sum_{i=1}^{n} (1 - z_i)\, (\overline{y}_1 - \overline{y}_0)$$

and similarly

$$\overline{x}_1 - \overline{x} = \frac{1}{n}\sum_{i=1}^{n} (1 - z_i)\, (\overline{x}_1 - \overline{x}_0)$$

and hence

$$\widehat{\boldsymbol{\beta}}_{\text{iv}} = \frac{\frac{1}{n}\sum_{i=1}^{n} (1 - z_i)\, (\overline{y}_1 - \overline{y}_0)}{\frac{1}{n}\sum_{i=1}^{n} (1 - z_i)\, (\overline{x}_1 - \overline{x}_0)} = \widehat{\beta}$$

as defined in (11.33). Thus the Wald estimator equals the IV estimator.

We can illustrate using the Card proximity example. If we estimate a simple IV model with no covariates we obtain the estimate $\widehat{\boldsymbol{\beta}}_{\text{iv}} = 0.19$. If we estimate the group-mean log wages and education levels based on the instrument *college*, we find

	near college	not near college
log(wage)	6.311	6.156
education	13.527	12.698

Based on these estimates the Wald estimator of the slope coefficient is $(6.311 - 6.156) \,/\, (13.527 - 12.698) = 0.19$, the same as the IV estimator.

11.11 Two-Stage Least Squares

The IV estimator described in the previous section presumed $\ell = k$. Now we allow the general case of $\ell \geq k$. Examining the reduced-form equation (11.16) we see

$$\begin{aligned} y_i &= \boldsymbol{z}_i' \boldsymbol{\Gamma} \boldsymbol{\beta} + v_i \\ \mathbb{E}\left(\boldsymbol{z}_i v_i\right) &= \mathbf{0}. \end{aligned}$$

Defining $\boldsymbol{w}_i = \boldsymbol{\Gamma}' \boldsymbol{z}_i$ we can write this as

$$\begin{aligned} y_i &= \boldsymbol{w}_i' \boldsymbol{\beta} + v_i \\ \mathbb{E}\left(\boldsymbol{w}_i v_i\right) &= \mathbf{0}. \end{aligned}$$

Suppose that $\boldsymbol{\Gamma}$ were known. Then we would estimate $\boldsymbol{\beta}$ by least-squares of y_i on $\boldsymbol{w}_i = \boldsymbol{\Gamma}' \boldsymbol{z}_i$

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{W}'\boldsymbol{W}\right)^{-1} \left(\boldsymbol{W}'\boldsymbol{y}\right) \\ &= \left(\boldsymbol{\Gamma}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\Gamma}\right)^{-1} \left(\boldsymbol{\Gamma}'\boldsymbol{Z}'\boldsymbol{y}\right). \end{aligned}$$

While this is infeasible, we can estimate $\boldsymbol{\Gamma}$ from the reduced form regression. Replacing $\boldsymbol{\Gamma}$ with its estimate $\widehat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \left(\boldsymbol{Z}'\boldsymbol{X}\right)$ we obtain

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{2sls}} &= \left(\widehat{\boldsymbol{\Gamma}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{\Gamma}}\right)^{-1} \left(\widehat{\boldsymbol{\Gamma}}'\boldsymbol{Z}'\boldsymbol{y}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}. \end{aligned} \tag{11.34}$$

This is called the **two-stage-least squares** (2SLS) estimator. It was originally proposed by Theil (1953) and Basmann (1957), and is a standard estimator for linear equations with instruments.

If the model is just-identified, so that $k = \ell$, then 2SLS simplifies to the IV estimator of the previous section. Since the matrices $\boldsymbol{X}'\boldsymbol{Z}$ and $\boldsymbol{Z}'\boldsymbol{X}$ are square, we can factor

$$\begin{aligned}\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1} \\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}.\end{aligned}$$

(Once again, this only works when $k = \ell$.) Then

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{2\text{sls}} &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} \\ &= \widehat{\boldsymbol{\beta}}_{\text{iv}}\end{aligned}$$

as claimed. This shows that the 2SLS estimator as defined in (11.34) is a generalization of the IV estimator defined in (11.29).

There are several alternative representations of the 2SLS estimator which we now describe. First, defining the projection matrix

$$\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}' \tag{11.35}$$

we can write the 2SLS estimator more compactly as

$$\widehat{\boldsymbol{\beta}}_{2\text{sls}} = \left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y}. \tag{11.36}$$

This is useful for representation and derivations, but is not useful for computation as the $n \times n$ matrix $\boldsymbol{P}_{\boldsymbol{Z}}$ is too large to compute when n is large.

Second, define the fitted values for $\boldsymbol{X}$ from the reduced form

$$\widehat{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X} = \boldsymbol{Z}\widehat{\boldsymbol{\Gamma}}.$$

Then the 2SLS estimator can be written as

$$\widehat{\boldsymbol{\beta}}_{2\text{sls}} = \left(\widehat{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{y}.$$

This is an IV estimator as defined in the previous section using $\widehat{\boldsymbol{X}}$ as the instrument.

Third, since $\boldsymbol{P}_{\boldsymbol{Z}}$ is idempotent, we can also write the 2SLS estimator as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{2\text{sls}} &= \left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y} \\ &= \left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{y}\end{aligned}$$

which is the least-squares estimator obtained by regressing $\boldsymbol{y}$ on the fitted values $\widehat{\boldsymbol{X}}$.

This is the source of the "two-stage" name is since it can be computed as follows.

- First regress $\boldsymbol{X}$ on $\boldsymbol{Z}$, vis., $\widehat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$ and $\widehat{\boldsymbol{X}} = \boldsymbol{Z}\widehat{\boldsymbol{\Gamma}} = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$.
- Second, regress $\boldsymbol{y}$ on $\widehat{\boldsymbol{X}}$, vis., $\widehat{\boldsymbol{\beta}}_{2\text{sls}} = \left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{y}$.

It is useful to scrutinize the projection $\widehat{\boldsymbol{X}}$. Recall, $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $\boldsymbol{Z} = [\boldsymbol{X}_1, \boldsymbol{Z}_2]$. Notice $\widehat{\boldsymbol{X}}_1 = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_1 = \boldsymbol{X}_1$ since $\boldsymbol{X}_1$ lies in the span of $\boldsymbol{Z}$. Then

$$\widehat{\boldsymbol{X}} = \left[\widehat{\boldsymbol{X}}_1, \widehat{\boldsymbol{X}}_2\right] = \left[\boldsymbol{X}_1, \widehat{\boldsymbol{X}}_2\right].$$

Thus in the second stage, we regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$ and $\widehat{\boldsymbol{X}}_2$. So only the endogenous variables $\boldsymbol{X}_2$ are replaced by their fitted values:

$$\widehat{\boldsymbol{X}}_2 = \boldsymbol{X}_1\widehat{\boldsymbol{\Gamma}}_{12} + \boldsymbol{Z}_2\widehat{\boldsymbol{\Gamma}}_{22}.$$

This least squares estimator can be written as

$$\boldsymbol{y} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \widehat{\boldsymbol{X}}_2\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{\varepsilon}}.$$

A fourth representation of 2SLS can be obtained from the previous representation for $\widehat{\boldsymbol{\beta}}_2$. Set $\boldsymbol{P}_1 = \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$. Applying the FWL theorem we obtain

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_2 &= \left(\widehat{\boldsymbol{X}}_2'\left(\boldsymbol{I}_n - \boldsymbol{P}_1\right)\widehat{\boldsymbol{X}}_2\right)^{-1}\left(\widehat{\boldsymbol{X}}_2'\left(\boldsymbol{I}_n - \boldsymbol{P}_1\right)\boldsymbol{y}\right) \\
&= \left(\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\left(\boldsymbol{I}_n - \boldsymbol{P}_1\right)\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\left(\boldsymbol{I}_n - \boldsymbol{P}_1\right)\boldsymbol{y}\right) \\
&= \left(\boldsymbol{X}_2'\left(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1\right)\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\left(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1\right)\boldsymbol{y}\right)
\end{aligned}$$

since $\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{P}_1 = \boldsymbol{P}_1$.

A fifth representation can be obtained by a further projection. The projection matrix $\boldsymbol{P}_{\boldsymbol{Z}}$ can be replaced by the projection onto the pair $[\boldsymbol{X}_1, \widetilde{\boldsymbol{Z}}_2]$ where $\widetilde{\boldsymbol{Z}}_2 = \left(\boldsymbol{I}_n - \boldsymbol{P}_1\right)\boldsymbol{Z}_2$ is $\boldsymbol{Z}_2$ projected orthogonal to $\boldsymbol{X}_1$. Since $\boldsymbol{X}_1$ and $\widetilde{\boldsymbol{Z}}_2$ are orthogonal, $\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{P}_1 + \boldsymbol{P}_2$ where $\boldsymbol{P}_2 = \widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'$. Thus $\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1 = \boldsymbol{P}_2$ and

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_2 &= \left(\boldsymbol{X}_2'\boldsymbol{P}_2\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{P}_2\boldsymbol{y}\right) \\
&= \left(\boldsymbol{X}_2'\widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{y}\right) \qquad (11.37)
\end{aligned}$$

Given the 2SLS estimator we define the residual vector

$$\widehat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{2\text{sls}}.$$

When the model is overidentified, the instruments and residuals are not orthogonal. That is

$$\boldsymbol{Z}'\widehat{\boldsymbol{e}} \neq \boldsymbol{0}.$$

It does, however, satisfy

$$\begin{aligned}
\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{e}} &= \widehat{\boldsymbol{\Gamma}}'\boldsymbol{Z}'\widehat{\boldsymbol{e}} \\
&= \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}} \\
&= \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{2\text{sls}} \\
&= \boldsymbol{0}.
\end{aligned}$$

Returning to Card's college proxity example, suppose that we treat experience as exogeneous, but that instead of using the single instrument *college* (grew up near a 4-year college) we use the two instruments (*public, private*) (grew up near a public/private 4-year college, respectively). In this case we have one endogenous variable (*education*) and two instruments (*public, private*). The estimated reduced form equation for *education* is presented in the sixth column of Table 11.2. In this specification, the coefficient on *public* – growing up near a public 4-year college – is larger

than that found for the variable *college* in the previous specification (column 2). Furthermore, the coefficient on *private* – growing up near a private 4-year college – is much smaller. This indicates that the key impact of proximity on education is via public colleges rather than private colleges.

The 2SLS estimates obtained using these two instruments are presented in the fourth column of Table 11.1. The coefficient on *education* increases to 0.162, indicating a 16% return to a year of education. This is roughly twice as large as the estimate obtained by least-squares in the first column.

Additionally, if we follow Card and treat *experience* as endogenous and use *age* as an instrument, we now have three endogenous variables (*education*, *experience, experience*$^2/100$) and four instruments (*public, private, age, age*^{2}). We present the 2SLS estimates using this specification in the fifth column of Table 11.1. The estimate of the return to education remains about 16%, but again the return to experience flattens.

You might wonder if we could use all three instruments – *college*, *public*, and *private*. The answer is no. This is because *college* = *public* + *private* so the three variables are colinear. Since the instruments are linearly related, the three together would violate the full-rank condition (11.7).

The 2SLS estimator may be calculated in Stata using the `ivregress 2sls` command.

11.12 Limited Information Maximum Likeihood

An alternative method to estimate the parameters of the structural equation is by maximum likelihood. Anderson and Rubin (1949) derived the maximum likelihood estimator for the joint distribution of $(y_i, \boldsymbol{x}_{2i})$. The estimator is known as **limited information maximum likelihood**, or LIML.

This estimator is called "limited information" because it is based on the structural equation for y_i combined with the reduced form equation for $\boldsymbol{x}_{2i}$. If maximum likelihood is derived based on a structural equation for $\boldsymbol{x}_{2i}$ as well, then this leads to what is known as **full information maximum likelihood** (FIML). The advantage of the LIML approach relative to FIML is that the former does not require a structural model for $\boldsymbol{x}_{2i}$, and thus allows the researcher to focus on the structural equation of interest – that for y_i. We do not describe the FIML estimator here as it is not commonly used in applied econometric practice.

While the LIML estimator is less widely used among economists than 2SLS, it has received a resurgence of attention from econometric theorists.

To derive the LIML estimator, start by writing the joint reduced form equations (11.19) and (11.15) as

$$\begin{aligned}\boldsymbol{w}_i &= \begin{pmatrix} y_i \\ \boldsymbol{x}_{2i} \end{pmatrix} \\ &= \begin{bmatrix} \boldsymbol{\lambda}_1' & \boldsymbol{\lambda}_2' \\ \boldsymbol{\Gamma}_{12}' & \boldsymbol{\Gamma}_{22}' \end{bmatrix} \begin{pmatrix} \boldsymbol{z}_{1i} \\ \boldsymbol{z}_{2i} \end{pmatrix} + \begin{pmatrix} v_i \\ \boldsymbol{u}_{2i} \end{pmatrix} \\ &= \boldsymbol{\Pi}_1' \boldsymbol{z}_{1i} + \boldsymbol{\Pi}_2' \boldsymbol{z}_{2i} + \boldsymbol{\xi}_i \end{aligned} \tag{11.38}$$

where $\boldsymbol{\Pi}_1 = \begin{bmatrix} \boldsymbol{\lambda}_1 & \boldsymbol{\Gamma}_{12} \end{bmatrix}$, $\boldsymbol{\Pi}_2 = \begin{bmatrix} \boldsymbol{\lambda}_2 & \boldsymbol{\Gamma}_{22} \end{bmatrix}$and $\boldsymbol{\xi}_i' = \begin{bmatrix} v_i & \boldsymbol{u}_{2i}' \end{bmatrix}$. The LIML estimator is derived under the assumption that $\boldsymbol{\xi}_i$ is multivariate normal.

Define $\boldsymbol{\gamma}' = \begin{bmatrix} 1 & -\boldsymbol{\beta}_2' \end{bmatrix}$. From (11.21) we find

$$\boldsymbol{\Pi}_2 \boldsymbol{\gamma} = \boldsymbol{\lambda}_2 - \boldsymbol{\Gamma}_{22} \boldsymbol{\beta}_2 = \mathbf{0}.$$

Thus the $\ell_2 \times (k_2 + 1)$ coefficient matrix $\boldsymbol{\Pi}_2$ in (11.38) has deficient rank. Indeed, its rank must be k_2, since $\boldsymbol{\Gamma}_{22}$ has full rank.

This means that the model (11.38) is precisely the reduced rank regression model of Section 10.9. Theorem 10.9.1 presents the maximum likelihood estimators for the reduced rank parameters.

In particular, the MLE for $\boldsymbol{\gamma}$ is

$$\widehat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \frac{\boldsymbol{\gamma}'\boldsymbol{W}'\boldsymbol{M}_1\boldsymbol{W}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{W}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{W}\boldsymbol{\gamma}} \tag{11.39}$$

where $\boldsymbol{W}$ is the $n \times (1+k_2)$ matrix of the stacked $\boldsymbol{w}_i' = \begin{pmatrix} y_i & \boldsymbol{x}_{2i}' \end{pmatrix}$, $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{Z}_1 \left(\boldsymbol{Z}_1'\boldsymbol{Z}_1\right)^{-1} \boldsymbol{Z}_1'$ and $\boldsymbol{M}_{\boldsymbol{Z}} = \boldsymbol{I}_n - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$. The minimization (11.39) is sometimes called the "least variance ratio" problem.

The minimization problem (11.39) is invariant to the scale of γ (that is, $\widehat{\boldsymbol{\gamma}}c$ is equivalently the argmin for any c) so a normalization is required. For estimation of the structural parameters a convenient normalization is $\boldsymbol{\gamma}' = \begin{bmatrix} 1 & -\boldsymbol{\beta}_2' \end{bmatrix}$. Another is to set $\boldsymbol{\gamma}'\boldsymbol{W}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{W}\boldsymbol{\gamma} = 1$. In this case, from the theory of the minimum of quadratic forms (Section A.11), $\widehat{\boldsymbol{\gamma}}$ is the generalized eigenvector of $\boldsymbol{W}'\boldsymbol{M}_1\boldsymbol{W}$ with respect to $\boldsymbol{W}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{W}$ associated with the smalled generalized eigenvalue. (See Section A.10 for the definition of generalized eigenvalues and eigenvectors.) Computationally this is straightforward. For example, in MATLAB, the generalized eigenvalues and eigenvectors of the matrix $\boldsymbol{A}$ with respect to $\boldsymbol{B}$ is found by the command `eig(A,B)`. Once $\widehat{\boldsymbol{\gamma}}$ is found, to obtain the MLE for $\boldsymbol{\beta}_2$, make the partition $\widehat{\boldsymbol{\gamma}}' = \begin{bmatrix} \widehat{\gamma}_1 & \widehat{\boldsymbol{\gamma}}_2' \end{bmatrix}$ and set $\widehat{\boldsymbol{\beta}}_2 = -\widehat{\boldsymbol{\gamma}}_2/\widehat{\gamma}_1$.

To obtain the MLE for $\boldsymbol{\beta}_1$, recall the structural equation $y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i$. Replacing $\boldsymbol{\beta}_2$ with the MLE $\widehat{\boldsymbol{\beta}}_2$ and then applying regression we obtain the MLE for $\boldsymbol{\beta}_1$. Thus

$$\widehat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right). \tag{11.40}$$

These solutions are the MLE (known as the LIML estimator) for the structural parameters $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$.

Many previous econometrics textbooks do not present a derivation of the LIML estimator as the original derivation by Anderson and Rubin (1949) is lengthy and not particularly insightful. In contrast, the derivation given here based on reduced rank regression is relatively simple.

There is an alternative (and traditional) expression for the LIML estimator. Define the minimum obtained in (11.39)

$$\widehat{\kappa} = \min_{\boldsymbol{\gamma}} \frac{\boldsymbol{\gamma}'\boldsymbol{W}'\boldsymbol{M}_1\boldsymbol{W}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{W}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{W}\boldsymbol{\gamma}} \tag{11.41}$$

which is the smallest generalized eigenvalue of $\boldsymbol{W}'\boldsymbol{M}_1\boldsymbol{W}$ with respect to $\boldsymbol{W}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{W}$. The LIML estimator then can be written as

$$\widehat{\boldsymbol{\beta}}_{\text{liml}} = \left(\boldsymbol{X}'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{y}\right). \tag{11.42}$$

We defer the derivation of (11.42) until the end of this section. Expression (11.42) does not simplify the computation (since $\widehat{\kappa}$ requires solving the same eigenvector problem that yields $\widehat{\boldsymbol{\beta}}_2$). However (11.42) is important for the distribution theory of of the LIML estimator, and to reveal the algebraic connection between LIML, least-squares, and 2SLS.

The estimator class (11.42) with arbitrary κ is known as a k class estimator of $\boldsymbol{\beta}$. While the LIML estimator obtains by setting $\kappa = \widehat{\kappa}$, the least-squares estimator is obtained by setting $\kappa = 0$ and 2SLS is obtained by setting $\kappa = 1$. It is worth observing that the LIML solution to (11.41) satisfies $\widehat{\kappa} \geq 1$.

When the model is just-identified, the LIML estimator is identical to the IV and 2SLS estimators. They are only different in the over-identified setting. (One corollary is that under just-identification the IV estimator is MLE under normality.)

For inference, it is useful to observe that (11.42) shows that $\widehat{\boldsymbol{\beta}}_{\text{liml}}$ can be written as an IV estimator

$$\widehat{\boldsymbol{\beta}}_{\text{liml}} = \left(\widetilde{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\left(\widetilde{\boldsymbol{X}}'\boldsymbol{y}\right)$$

using the instrument

$$\widetilde{\boldsymbol{X}} = \left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 - \widehat{\kappa}\widehat{\boldsymbol{U}}_2 \end{pmatrix}$$

where $\widehat{\boldsymbol{U}}_2 = \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}_2$ are the (reduced-form) residuals from the multivariate regression of the endogenous regressors $\boldsymbol{x}_{2i}$ on the instruments $\boldsymbol{z}_i$. Expressing LIML using this IV formula is useful for variance estimation.

Asymptotically the LIML estimator has the same distribution as 2SLS. However, they can have quite different behaviors in finite samples. There is considerable evidence that the LIML estimator has superior finite sample performance to 2SLS when there are many instruments or the reduced form is weak. (We review these cases in the following sections.) However, on the other hand there is worry that since the LIML estimator is derived under normality it may not be robust in non-normal settings.

We now derive the expression (11.42). Use the normaliaation $\boldsymbol{\gamma}' = \begin{bmatrix} 1 & -\boldsymbol{\beta}_2' \end{bmatrix}$ to write (11.39) as

$$\widehat{\boldsymbol{\beta}}_2 = \underset{\boldsymbol{\beta}_2}{\operatorname{argmin}} \frac{(\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2)' \boldsymbol{M}_1 (\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2)}{(\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2)' \boldsymbol{M}_{\boldsymbol{Z}} (\boldsymbol{Y} - \boldsymbol{X}_2\boldsymbol{\beta}_2)}$$

The first-order-condition for minimization

$$2\frac{\boldsymbol{X}_2'\boldsymbol{M}_1\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)}{\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)} - 2\frac{\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)'\boldsymbol{M}_1\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)}{\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)^2}\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right) = \boldsymbol{0}.$$

Multiplying by $\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)/2$ and using definition (11.41) we find

$$\boldsymbol{X}_2'\boldsymbol{M}_1\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right) - \widehat{\kappa}\boldsymbol{X}_2'\boldsymbol{M}_{\boldsymbol{Z}}\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right) = \boldsymbol{0}.$$

Rewriting,

$$\boldsymbol{X}_2'\left(\boldsymbol{M}_1 - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2'\left(\boldsymbol{M}_1 - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{y}. \tag{11.43}$$

Equation (11.42) is the same as the two equation system

$$\begin{aligned} \boldsymbol{X}_1'\boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_1'\boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 &= \boldsymbol{X}_1'\boldsymbol{y} \\ \boldsymbol{X}_2'\boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \left(\boldsymbol{X}_2'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2\right)\widehat{\boldsymbol{\beta}}_2 &= \boldsymbol{X}_2'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{y}. \end{aligned}$$

The first equation is (11.40). Using (11.40), the second is

$$\boldsymbol{X}_2'\boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\left(\boldsymbol{Y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right) + \left(\boldsymbol{X}_2'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2\right)\widehat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_2'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{y}$$

which is (11.43) when rearranged. We have thus shown that (11.42) is equivalent to (11.40) and (11.43) and is thus a valid expression for the LIML estimator.

Returning to the Card college proximity example, we now present the LIML estimates of the equation with the two instruments (*public*, *private*). They are reported in the final column of Table 11.1. They are quite similar to the 2SLS estimates in this application.

The LIML estimator may be calculated in Stata using the `ivregress liml` command.

Theodore Anderson

Theodore (Ted) Anderson (1918-2016) was a American statistician and econometrician, who made fundamental contributions to multivariate statistical theory. Important contributions include the Anderson-Darling distribution test, the Anderson-Rubin statistic, the method of reduced rank regression, and his most famous econometrics contribution – the LIML estimator. He continued working throughout his long life, even publishing theoretical work at the age of 97!

11.13 Consistency of 2SLS

We now present a demonstration of the consistency of the 2SLS estimate for the structural parameter. The following is a set of regularity conditions.

Assumption 11.13.1

1. *The observations* $(y_i, \boldsymbol{x_i}, \boldsymbol{z_i})$, $i = 1, ..., n$, *are independent and identically distributed.*
2. $\mathbb{E}\left(y^2\right) < \infty$.
3. $\mathbb{E}\left\|\boldsymbol{x}\right\|^2 < \infty$.
4. $\mathbb{E}\left\|\boldsymbol{z}\right\|^2 < \infty$.
5. $\mathbb{E}\left(\boldsymbol{z}\boldsymbol{z}'\right)$ *is positive definite.*
6. $\mathbb{E}\left(\boldsymbol{z}\boldsymbol{x}'\right)$ *has full rank* k.
7. $\mathbb{E}\left(\boldsymbol{z}\boldsymbol{e}\right) = 0$.

Assumptions 11.13.1.2-4 state that all variables have finite variances. Assumption 11.13.1.5 states that the instrument vector has an invertible design matrix, which is identical to the core assumption about regressors in the linear regression model. This excludes linearly redundant instruments. Assumptions 11.13.1.6 and 11.13.1.7 are the key identification conditions for instrumental variables. Assumption 11.13.1.6 states that the instruments and regressors have a full-rank cross-moment matrix. This is often called the relevance condition. Assumption 11.13.1.7 states that the instrumental variables and structural error are uncorrelated. Assumptions 11.13.1.5-7 are identical to Definition 11.3.1.

Theorem 11.13.1 *Under Assumption 11.13.1,* $\widehat{\boldsymbol{\beta}}_{2\text{sls}} \xrightarrow{p} \boldsymbol{\beta}$ *as* $n \to \infty$.

The proof of the theorem is provided below

This theorem shows that the 2SLS estimator is consistent for the structural coefficient $\boldsymbol{\beta}$ under similar moment conditions as the least-squares estimator. The key differences are the instrumental variables assumption $\mathbb{E}\left(\boldsymbol{z}\boldsymbol{e}\right) = 0$ and the identification assumption $\operatorname{rank}\left(\mathbb{E}\left(\boldsymbol{z}\boldsymbol{x}'\right)\right) = k$.

The result includes the IV estimator (when $\ell = k$) as a special case.

The proof of this consistency result is similar to that for the least-squares estimator. Take the structural equation $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ in matrix format and substitute it into the expression for the estimator. We obtain

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{2\text{sls}} &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}\right) \\ &= \boldsymbol{\beta} + \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{e}. \qquad (11.44)\end{aligned}$$

This separates out the stochastic component. Re-writing and applying the WLLN and CMT

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{2\text{sls}} - \boldsymbol{\beta} &= \left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1} \\ &\cdot \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{e}\right) \\ &\xrightarrow{p} \left(\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\boldsymbol{Q}_{\boldsymbol{zx}}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\mathbb{E}\left(\boldsymbol{z}_i e_i\right) = 0\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{Q}_{\boldsymbol{xz}} &= \mathbb{E}\left(\boldsymbol{x}_{\boldsymbol{i}}\boldsymbol{z}_i'\right) \\ \boldsymbol{Q}_{\boldsymbol{zz}} &= \mathbb{E}\left(\boldsymbol{z}_{\boldsymbol{i}}\boldsymbol{z}_i'\right) \\ \boldsymbol{Q}_{\boldsymbol{zx}} &= \mathbb{E}\left(\boldsymbol{z}_{\boldsymbol{i}}\boldsymbol{x}_i'\right).\end{aligned}$$

The WLLN holds under the i.i.d. Assumption 11.13.1.1 and the finite second moment Assumptions 11.13.1.2-4. The continuous mapping theorem applies if the matrices $\boldsymbol{Q}_{\boldsymbol{zz}}$ and $\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\boldsymbol{Q}_{\boldsymbol{zx}}$ are invertible, which hold under the identification Assumptions 11.13.1.5 and 11.13.1.6. The final equality uses Assumption 11.13.1.7.

11.14 Asymptotic Distribution of 2SLS

We now show that the 2SLS estimator satisfies a central limit theorem. We first state a set of sufficient regularity conditions.

Assumption 11.14.1 *In addition to Assumption 11.13.1,*

1. $\mathbb{E}\left(y^4\right) < \infty$.
2. $\mathbb{E}\left\|\boldsymbol{z}\right\|^4 < \infty$.

Assumption 11.14.1 strengthens Assumption 11.13.1 by requiring that the dependent variable and instruments have finite fourth moments. This is used to establish the central limit theorem.

Theorem 11.14.1 *Under Assumption 11.14.1, as $n \to \infty$.*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{2\text{sls}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\boldsymbol{Q}_{\boldsymbol{zx}}\right)^{-1}\left(\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\boldsymbol{Q}_{\boldsymbol{zx}}\right)\left(\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\boldsymbol{Q}_{\boldsymbol{zx}}\right)^{-1}$$

and

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i' e_i^2\right).$$

This shows that the 2SLS estimator converges at a $\sqrt{n}$ rate to a normal random vector. It shows as well the form of the covariance matrix. The latter takes a substantially more complicated form than the least-squares estimator.

As in the case of least-squares estimation, the asymptotic variance simplifies under a conditional homoskedasticity condition. For 2SLS the simplification occurs when $\mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right) = \sigma^2$. This holds when $\boldsymbol{z}_i$ and e_i are independent. It may be reasonable in some contexts to conceive that the error e_i is independent of the excluded instruments $\boldsymbol{z}_{2i}$, since by assumption the impact of $\boldsymbol{z}_{2i}$ on y_i is only through $\boldsymbol{x}_i$, but there is no reason to expect e_i to be independent of the included exogenous variables $\boldsymbol{x}_{1i}$. Hence heteroskedasticity should be equally expected in 2SLS and least-squares regression. Nevertheless, under the homoskedasticity condition then we have the simplifications $\boldsymbol{\Omega} = \boldsymbol{Q}_{\boldsymbol{zz}}\sigma^2$ and $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\beta}}^0 \stackrel{def}{=} \left(\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\boldsymbol{Q}_{\boldsymbol{zx}}\right)^{-1}\sigma^2$.

The derivation of the asymptotic distribution builds on the proof of consistency. Using equation (11.44) we have

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{2\text{sls}} - \boldsymbol{\beta}\right) = &\left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1} \\ &\cdot\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\boldsymbol{e}\right).\end{aligned}$$

We apply the WLLN and CMT for the moment matrices involving $\boldsymbol{X}$ and $\boldsymbol{Z}$ the same as in the proof of consistency. In addition, by the CLT for i.i.d. observations

$$\frac{1}{\sqrt{n}}\boldsymbol{Z}'\boldsymbol{e} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{z}_i e_i \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right)$$

because the vector $\boldsymbol{z}_i e_i$ is i.i.d. and mean zero under Assumptions 11.13.1.1 and 11.13.1.7, and has a finite second moment as we verify below.

We obtain

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{2\text{sls}} - \boldsymbol{\beta}\right) &= \left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1} \\ &\quad\cdot\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\boldsymbol{e}\right) \\ &\xrightarrow{d} \left(\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\boldsymbol{Q}_{\boldsymbol{zx}}\right)^{-1}\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right) = \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)\end{aligned}$$

as stated.

For completeness, we demonstrate that $\boldsymbol{z}_i e_i$ has a finite second moment under Assumption 11.14.1. To see this, note that by Minkowski's inequality

$$\begin{aligned}\left(\mathbb{E}\left(e^4\right)\right)^{1/4} &= \left(\mathbb{E}\left(\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)^4\right)\right)^{1/4} \\ &\le \left(\mathbb{E}\left(y^4\right)\right)^{1/4} + \|\boldsymbol{\beta}\|\left(\mathbb{E}\left\|\boldsymbol{x}\right\|^4\right)^{1/4} < \infty\end{aligned}$$

under Assumptions 11.14.1.1 and 11.14.1.2. Then by the Cauchy-Schwarz inequality

$$\mathbb{E}\left\|\boldsymbol{z}e\right\|^2 \le \left(\mathbb{E}\left\|\boldsymbol{z}\right\|^4\right)^{1/2}\left(\mathbb{E}\left(e^4\right)\right)^{1/2} < \infty$$

using Assumptions 11.14.1.3.

11.15 Determinants of 2SLS Variance

It is instructive to examine the asymptotic variance of the 2SLS estimator to understand the factors which determine the precision (or lack thereof) of the estimator. As in the least-squares case, it is more transparent to examine the variance under the assumption of homoskedasticity. In this case the asymptotic variance takes the form

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{\beta}}^{0} &= \left(\boldsymbol{Q}_{\boldsymbol{xz}}\boldsymbol{Q}_{\boldsymbol{zz}}^{-1}\boldsymbol{Q}_{\boldsymbol{zx}}\right)^{-1}\sigma^{2} \\ &= \left(\mathbb{E}\left(\boldsymbol{x}_{i}\boldsymbol{z}_{i}'\right)\left(\mathbb{E}\left(\boldsymbol{z}_{i}\boldsymbol{z}_{i}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{i}\boldsymbol{x}_{i}'\right)\right)^{-1}\mathbb{E}\left(e_{i}^{2}\right).\end{aligned}$$

As in the least-squares case, we can see that the variance is increasing in the variance of the error e_i, and decreasing in the variance of $\boldsymbol{x}_i$. What is different is that the variance is decreasing in the (matrix-valued) correlation between $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$.

It is also useful to observe that the variance expression is not affected by the variance structure of $\boldsymbol{z}_i$. Indeed, $\boldsymbol{V}_{\boldsymbol{\beta}}^{0}$ is invariant to rotations of $\boldsymbol{z}_i$ (if you replace $\boldsymbol{z}_i$ with $\boldsymbol{C}\boldsymbol{z}_i$ for invertible $\boldsymbol{C}$ the expression does not change). This means that the variance expression is not affected by the scaling of $\boldsymbol{z}_i$, and is not directly affected by correlation among the $\boldsymbol{z}_i$.

We can also use this expression to examine the impact of increasing the instrument set. Suppose we partition $\boldsymbol{z}_i = (\boldsymbol{z}_{ai}, \boldsymbol{z}_{bi})$ where $\dim(\boldsymbol{z}_{ai}) \geq k$ so we can construct the 2SLS estimator using $\boldsymbol{z}_{ai}$. Let $\widehat{\boldsymbol{\beta}}_a$ and $\widehat{\boldsymbol{\beta}}$ denote the 2SLS estimators constructed using the instrument sets $\boldsymbol{z}_{ai}$ and $(\boldsymbol{z}_{ai}, \boldsymbol{z}_{bi})$, respectively. Without loss of generality we can assume that $\boldsymbol{z}_{ai}$ and $\boldsymbol{z}_{bi}$ are uncorrelated (if not, replace $\boldsymbol{z}_{bi}$ with the projection error after projecting onto $\boldsymbol{z}_{ai}$). In this case both $\mathbb{E}\left(\boldsymbol{z}_{i}\boldsymbol{z}_{i}'\right)$ and $\left(\mathbb{E}\left(\boldsymbol{z}_{i}\boldsymbol{z}_{i}'\right)\right)^{-1}$ are block diagonal, so

$$\begin{aligned}\operatorname{avar}\left(\widehat{\boldsymbol{\beta}}\right) &= \left(\mathbb{E}\left(\boldsymbol{x}_{i}\boldsymbol{z}_{i}'\right)\left(\mathbb{E}\left(\boldsymbol{z}_{i}\boldsymbol{z}_{i}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{i}\boldsymbol{x}_{i}'\right)\right)^{-1}\sigma^{2} \\ &= \left(\mathbb{E}\left(\boldsymbol{x}_{i}\boldsymbol{z}_{ai}'\right)\left(\mathbb{E}\left(\boldsymbol{z}_{ai}\boldsymbol{z}_{ai}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{ai}\boldsymbol{x}_{i}'\right) + \mathbb{E}\left(\boldsymbol{x}_{i}\boldsymbol{z}_{bi}'\right)\left(\mathbb{E}\left(\boldsymbol{z}_{bi}\boldsymbol{z}_{bi}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{bi}\boldsymbol{x}_{i}'\right)\right)^{-1}\sigma^{2} \\ &\leq \left(\mathbb{E}\left(\boldsymbol{x}_{i}\boldsymbol{z}_{ai}'\right)\left(\mathbb{E}\left(\boldsymbol{z}_{ai}\boldsymbol{z}_{ai}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{ai}\boldsymbol{x}_{i}'\right)\right)^{-1}\sigma^{2} \\ &= \operatorname{avar}\left(\widehat{\boldsymbol{\beta}}_{a}\right)\end{aligned}$$

with strict inequality if $\mathbb{E}\left(\boldsymbol{x}_{i}\boldsymbol{z}_{bi}'\right) \neq \boldsymbol{0}$. Thus the 2SLS estimator with the full instrument set has a smaller asymptotic variance than the estimator with the smaller instrument set.

What we have shown is that the asymptotic variance of the 2SLS estimator is decreasing as the number of instruments increases. From the viewpoint of asymptotic efficiency, thie means that it is better to use more instruments (when they are available and are all known to be valid instruments) rather than less.

Unfortunately, there is always a catch. In this case it turns out that the finite sample bias of the 2SLS estimator (which cannot be calculated exactly, but can be approximated using asymptotic expansions) is generically increasing linearily as the number of instruments increases. We will see some calculations illustrating this phenomenon in Section 11.33. Thus the choice of instruments in practice induces a trade-off between bias and variance.

11.16 Covariance Matrix Estimation

Estimation of the asymptotic variance matrix $\boldsymbol{V}_{\boldsymbol{\beta}}$ is done using similar techniques as for least-squares estimation. The estimator is constructed by replacing the population moment matrices by sample counterparts. Thus

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{Q}}_{\boldsymbol{xz}}\widehat{\boldsymbol{Q}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{Q}}_{\boldsymbol{zx}}\right)^{-1}\left(\widehat{\boldsymbol{Q}}_{\boldsymbol{xz}}\widehat{\boldsymbol{Q}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{Q}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{Q}}_{\boldsymbol{zx}}\right)\left(\widehat{\boldsymbol{Q}}_{\boldsymbol{xz}}\widehat{\boldsymbol{Q}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{Q}}_{\boldsymbol{zx}}\right)^{-1} \tag{11.45}$$

where

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{zz}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i' = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}$$

$$\widehat{\boldsymbol{Q}}_{\boldsymbol{xz}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{z}_i' = \frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}$$

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i' \widehat{e}_i^2$$

$$\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{2\text{sls}}.$$

The homoskedastic variance matrix can be estimated by

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 = \left(\widehat{\boldsymbol{Q}}_{\boldsymbol{xz}}\widehat{\boldsymbol{Q}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{Q}}_{\boldsymbol{zx}}\right)^{-1}\widehat{\sigma}^2$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2.$$

Standard errors for the coefficients are obtained as the square roots of the diagonal elements of $n^{-1}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$. Confidence intervals, t-tests, and Wald tests may all be constructed from the coefficient estimates and covariance matrix estimate exactly as for least-squares regression.

In Stata, the `ivregress` command by default calculates the covariance matrix estimator using the homoskedastic variance matrix. To obtain covariance matrix estimation and standard errors with the robust estimator $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$, use the "`,r`" option.

Theorem 11.16.1 *Under Assumption 11.14.1, as $n \to \infty$,*

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}^0$$

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}.$$

To prove Theorem 11.16.1 the key is to show $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$ as the other convergence results were established in the proof of consistency. We defer this to Exercise 11.6.

It is important that the covariance matrix be constructed using the correct residual formula $\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{2\text{sls}}$. This is different than what would be obtained if the "two-stage" computation method is used. To see this, let's walk through the two-stage method. First, we estimate the reduced form

$$\boldsymbol{x}_i = \widehat{\boldsymbol{\Gamma}}'\boldsymbol{z}_i + \widehat{\boldsymbol{u}}_i$$

to obtain the predicted values $\widehat{\boldsymbol{x}}_i = \widehat{\boldsymbol{\Gamma}}'\boldsymbol{z}_i$. Second, we regress y_i on $\widehat{\boldsymbol{x}}_i$ to obtain the 2SLS estimator $\widehat{\boldsymbol{\beta}}_{2\text{sls}}$. This latter regression takes the form

$$y_i = \widehat{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}_{2\text{sls}} + \widehat{v}_i \tag{11.46}$$

where $\widehat{v}_i$ are least-squares residuals. The covariance matrix (and standard errors) reported by this regression are constructed using the residual $\widehat{v}_i$. For example, the homoskedastic formula is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\frac{1}{n}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\sigma}_v^2 = \left(\widehat{\boldsymbol{Q}}_{\boldsymbol{xz}}\widehat{\boldsymbol{Q}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{Q}}_{\boldsymbol{zx}}\right)^{-1}\widehat{\sigma}_v^2$$

$$\widehat{\sigma}_v^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{v}_i^2$$

which is proportional to the variance estimate $\widehat{\sigma}_v^2$ rather than $\widehat{\sigma}^2$. This is important because the residual $\widehat{v}_i$ differs from $\widehat{e}_i$. We can see this because the regression (11.46) uses the regressor $\widehat{\boldsymbol{x}}_i$ rather than $\boldsymbol{x}_i$. Indeed, we can calculate that

$$\begin{aligned}\widehat{v}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{2\text{sls}} + (\boldsymbol{x}_i - \widehat{\boldsymbol{x}}_i)'\widehat{\boldsymbol{\beta}}_{2\text{sls}} \\ &= \widehat{e}_i + \widehat{\boldsymbol{u}}_i'\widehat{\boldsymbol{\beta}}_{2\text{sls}} \\ &\neq \widehat{e}_i\end{aligned}$$

This means that standard errors reported by the regression (11.46) will be incorrect.

This problem is avoided if the 2SLS estimator is constructed directly and the standard errors calculated with the correct formula rather than taking the "two-step" shortcut.

11.17 Asymptotic Distribution and Covariance Estimation for LIML

Recall, the LIML estimator has several representations, including

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{liml}} &= \left(\boldsymbol{X}'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{I}_n - \widehat{\kappa}\boldsymbol{M}_{\boldsymbol{Z}}\right)\boldsymbol{y}\right). \\ &= \left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X} - \widehat{\mu}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y} - \widehat{\mu}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{y}\right)\end{aligned}$$

where $\widehat{\mu} = \widehat{\kappa} - 1$ and

$$\widehat{\kappa} = \min_{\boldsymbol{\gamma}} \frac{\boldsymbol{\gamma}'\boldsymbol{W}'\boldsymbol{M}_1\boldsymbol{W}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\boldsymbol{W}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{W}\boldsymbol{\gamma}}.$$

Using multivariate regression analysis, we can show that $\widehat{\kappa} \xrightarrow{p} 1$ and thus $\widehat{\mu} \xrightarrow{p} 0$. It follows that

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{liml}} - \boldsymbol{\beta}\right) &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X} - \widehat{\mu}\frac{1}{n}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{e} - \widehat{\mu}\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{e}\right) \\ &= \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X} - o_p(1)\right)^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{e} - o_p(1)\right) \\ &= \sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{2\text{sls}} - \boldsymbol{\beta}\right) + o_p(1)\end{aligned}$$

which means that LIML and 2SLS have the same asymptotic distribution. This holds under the same assumptions as for 2SLS, and in particular does not require normality of the errors.

Consequently, one method to obtain an asymptotically valid covariance estimate for LIML is to use the same formula as for 2SLS. However, this is not the best choice. Rather, consider the IV representation for LIML

$$\widehat{\boldsymbol{\beta}}_{\text{liml}} = \left(\widetilde{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\left(\widetilde{\boldsymbol{X}}'\boldsymbol{y}\right)$$

where

$$\widetilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 - \widehat{\kappa}\widehat{\boldsymbol{U}}_2 \end{pmatrix}$$

and $\widehat{\boldsymbol{U}}_2 = \boldsymbol{M}_{\boldsymbol{Z}}\boldsymbol{X}_2$. The asymptotic covariance matrix formula for an IV estimator is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\frac{1}{n}\widetilde{\boldsymbol{X}}'\boldsymbol{X}\right)^{-1}\widehat{\boldsymbol{\Omega}}\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{X}}\right)^{-1} \tag{11.47}$$

where

$$\begin{aligned}\widehat{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{x}}_i\widetilde{\boldsymbol{x}}_i\widehat{e}_i^2 \\ \widehat{e}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{liml}}.\end{aligned}$$

This simplifies to the 2SLS formula when $\widehat{\kappa} = 1$ but otherwise differs. The estimator (11.47) is a better choice than the 2SLS formula for covariance matrix estimation as it takes advantage of the LIML estimator structure.

11.18 Functions of Parameters

Given the distribution theory in Theorems 11.14.1 and 11.16.1 it is straightforward to derive the asymptotic distribution of smooth nonlinear functions of the coefficients.

Specifically, given a function $\boldsymbol{r}(\boldsymbol{\beta}) : \mathbb{R}^k \to \Theta \subset \mathbb{R}^q$ we define the parameter

$$\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$$

Given $\widehat{\boldsymbol{\beta}}_{2\text{sls}}$ a natural estimator of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}}_{2\text{sls}} = \boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}_{2\text{sls}}\right)$.

Consistency follows from Theorem 11.13.1 and the continuous mapping theorem.

Theorem 11.18.1 *Under Assumption 11.13.1, if $\boldsymbol{r}(\boldsymbol{\beta})$ is continuous at $\boldsymbol{\beta}$, then $\widehat{\boldsymbol{\theta}}_{2\text{sls}} \xrightarrow{p} \boldsymbol{\theta}$ as $n \to \infty$.*

If $\boldsymbol{r}(\boldsymbol{\beta})$ is differentiable then an estimator of the asymptotic covariance matrix for $\widehat{\boldsymbol{\theta}}$ is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}$$
$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\widehat{\boldsymbol{\beta}}_{2\text{sls}})'$$

We similarly define the homoskedastic variance estimator as

$$\widehat{\boldsymbol{V}}^0_{\boldsymbol{\theta}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}^0_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}.$$

The asymptotic distribution theory follows from Theorems 11.14.1 and 11.16.1, and the delta method.

Theorem 11.18.2 *Under Assumption 11.14.1, if $\boldsymbol{r}(\boldsymbol{\beta})$ is continuously differentiable at $\boldsymbol{\beta}$, then as $n \to \infty$*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{2\text{sls}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{R}' \boldsymbol{V}_{\boldsymbol{\beta}} \boldsymbol{R}$$
$$\boldsymbol{R} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\boldsymbol{\beta})'$$

and

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\theta}}.$$

When $q = 1$, a standard error for $\widehat{\boldsymbol{\theta}}_{2\text{sls}}$ is $s(\widehat{\boldsymbol{\theta}}_{2\text{sls}}) = \sqrt{n^{-1} \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}}$.

For example, let's take the parameter estimates from the fifth column of Table 11.1, which are the 2SLS estimates with three endogenous regressors and four excluded instruments. Suppose we are interested in the return to experience, which depends on the level of experience. The estimated return at *experience* $= 10$ is $0.0473 - 0.032 * 2 * 10/100 = 0.041$ and its standard error is 0.003. This implies a 4% increase in wages per year of experience and is precisely estimated. Or suppose we are interested in the level of experience at which the function maximizes. The estimate is $50 * 0.047/0.032 = 73$. This has a standard error of 249. The large standard error implies that the estimate (73 years of experience) is without precision and is thus uninformative.

11.19 Hypothesis Tests

As in the previous section, for a given function $\boldsymbol{r}(\boldsymbol{\beta}) : \mathbb{R}^k \to \Theta \subset \mathbb{R}^q$ we define the parameter $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$ and consider tests of hypotheses of the form

$$\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

against

$$\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0.$$

The Wald statistic for $\mathbb{H}_0$ is

$$W = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).$$

From Theorem 11.18.2 we deduce that W is asymptotically chi-square distributed. Let $G_q(u)$ denote the χ_q^2 distribution function.

Theorem 11.19.1 *Under Assumption 11.14.1, if $\boldsymbol{r}(\boldsymbol{\beta})$ is continuously differentiable at $\boldsymbol{\beta}$, and $\mathbb{H}_0$ holds, then as $n \to \infty$,*

$$W \xrightarrow{d} \chi_q^2.$$

For c satisfying $\alpha = 1 - G_q(c)$,

$$\Pr\left(W > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ if $W > c$" has asymptotic size α.

In linear regression we often report the F version of the Wald statistic (by dividing by degrees of freedom) and use the F distribution for inference, as this is justified in the normal sampling model. For 2SLS estimation, however, this is not done as there is no finite sample F justification for the F version of the Wald statistic.

To illustrate, once again let's take the parameter estimates from the fifth column of Table 11.1 and again consider the return to experience which is determined by the coefficients on *experience* and *experience*$^2/100$. Neither coefficient is statisticially signfiicant at the 5% level, so it is unclear from a casual look if the overall effect is statistically significant. We can assess this by testing the joint hypothesis that both coefficients are zero. The Wald statistic for this hypothesis is $W = 254$, which is highly significant with an asymptotic p-value of 0.0000. Thus by examining the joint test, in contrast to the individual tests, is quite clear that experience has a non-zero effect.

11.20 Finite Sample Theory

In Chapter 5 we reviewed the rich exact distribution available for the linear regression model under the assumption of normal innovations. There was a similarly rich literature in econometrics which developed a distribution theory for IV, 2SLS and LIML estimators. This theory is reviewed by Peter Phillips (1983), and much of the theory was developed by Peter Phillips in a series of papers in the 1970s and early 1980s.

This theory was developed under the assumption that the structural error vector $\boldsymbol{e}$ and reduced form error $\boldsymbol{u}_2$ are multivariate normally distributed. The challenge is that the IV estimators are non-linear functions of $\boldsymbol{u}_2$ and are thus non-normally distributed. Formulae for the exact distributions have been derived, but are unfortunately functions of model parameters and hence are not directly useful for finite sample inference.

One important implication of this literature is that it is quite clear that even in this optimal context of exact normal innovations, the finite sample distributions of the IV estimators are non-normal and the finite sample distributions of test statistics are not chi-squared. The normal and chi-squared approximations hold asymptotically, but there is no reason to expect these approximations to be accurate in finite samples.

11.21 Clustered Dependence

In Section 4.20 we introduced clustered dependence. We can also use the methods of clustered dependence for 2SLS estimation. Recall, the g^{th} cluster has the observations $\boldsymbol{y}_g = (y_{1g}, ..., y_{n_g g})'$, $\boldsymbol{X}_g = (\boldsymbol{x}_{1g}, ..., \boldsymbol{x}_{n_g g})'$, and $\boldsymbol{Z}_g = (\boldsymbol{z}_{1g}, ..., \boldsymbol{z}_{n_g g})'$. The structural equation for the g^{th} cluster can be written as the matrix system

$$\boldsymbol{y}_g = \boldsymbol{X}_g \boldsymbol{\beta} + \boldsymbol{e}_g.$$

Using this notation the centere 2SLS estimator can be written as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_{2\text{sls}} - \boldsymbol{\beta} &= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{e} \\
&= \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{e}_g\right).
\end{aligned}$$

The cluster-robust covariance matrix estimator for $\widehat{\boldsymbol{\beta}}_{2\text{sls}}$ thus takes the form

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\widehat{\boldsymbol{S}}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}$$

with

$$\widehat{\boldsymbol{S}} = \sum_{g=1}^{G}\boldsymbol{Z}_g'\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{Z}_g$$

and the clustered residuals

$$\widehat{\boldsymbol{e}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g\widehat{\boldsymbol{\beta}}_{2\text{sls}}.$$

The difference between the heteroskedasticity-robust estimator and the cluster-robust estimator is the covariance estimator $\widehat{\boldsymbol{S}}$.

11.22 Generated Regressors

The "two-stage" form of the 2SLS estimator is an example of what is called "estimation with generated regressors". We say a regressor is a **generated** if it is an estimate of an idealized regressor, or if it is a function of estimated parameters. Typically, a generated regressor $\widehat{\boldsymbol{w}}_i$ is an estimate of an unobserved ideal regressor $\boldsymbol{w}_i$. As an estimate, $\widehat{\boldsymbol{w}}_i$ is a function of the sample, not just observation i. Hence it is not "i.i.d." as it is dependent across observations, which invalidates the conventional regression assumptions. Consequently, the sampling distribution of regression estimates is affected. Unless this is incorporated into our inference methods, covariance matrix estimates and standard errors will be incorrect.

The econometric theory of generated regressors was developed by Pagan (1984) for linear models, and extended to non-linear models and more general two-step estimators by Pagan (1986). Here we focus on the linear model:

$$\begin{aligned}
y_i &= \boldsymbol{w}_i'\boldsymbol{\beta} + v_i \\
\boldsymbol{w}_i &= \boldsymbol{A}'\boldsymbol{z}_i \\
\mathbb{E}\left(\boldsymbol{z}_i v_i\right) &= \boldsymbol{0}.
\end{aligned} \tag{11.48}$$

The observables are $(y_i, \boldsymbol{z}_i)$. We also have an estimate $\widehat{\boldsymbol{A}}$ of $\boldsymbol{A}$.

Given $\widehat{\boldsymbol{A}}$ we construct the estimate $\widehat{\boldsymbol{w}}_i = \widehat{\boldsymbol{A}}'\boldsymbol{z}_i$ of $\boldsymbol{w}_i$, replace $\boldsymbol{w}_i$ in (11.48) with $\widehat{\boldsymbol{w}}_i$, and then estimate $\boldsymbol{\beta}$ by least-squares, resulting in the estimator

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \widehat{\boldsymbol{w}}_i \widehat{\boldsymbol{w}}_i'\right)^{-1} \left(\sum_{i=1}^{n} \widehat{\boldsymbol{w}}_i y_i\right). \tag{11.49}$$

The regressors $\widehat{\boldsymbol{w}}_i$ are called **generated regressors**. The properties of $\widehat{\boldsymbol{\beta}}$ are different than least-squares with i.i.d. observations, since the generated regressors are themselves estimates.

This framework includes the 2SLS estimator as well as other common estimators. The 2SLS model can be written as (11.48) by looking at the reduced form equation (11.16), with $\boldsymbol{w}_i = \boldsymbol{\Gamma}'\boldsymbol{z}_i$, $\boldsymbol{A} = \boldsymbol{\Gamma}$, and $\widehat{\boldsymbol{A}} = \widehat{\boldsymbol{\Gamma}}$ is (11.22).

The examples which motivated Pagan (1984) emerged from the macroeconomics literature, in particular the work of Barro (1977) which examined the impact of inflation expectations and expectation errors on economic output. For example, let π_i denote realized inflation and $\boldsymbol{z}_i$ be the information available to economic agents. A model of inflation expectations sets $w_i = \mathbb{E}(\pi_i|\boldsymbol{z}_i) = \boldsymbol{\gamma}'\boldsymbol{z}_i$ and a model of expectation error sets $w_i = \pi_i - \mathbb{E}(\pi_i|\boldsymbol{z}_i) = \pi_i - \boldsymbol{\gamma}'\boldsymbol{z}_i$. Since expectations and errors are not observed they are replaced in applications with the fitted values $\widehat{w}_i = \widehat{\boldsymbol{\gamma}}'\boldsymbol{z}_i$ or residuals $\widehat{w}_i = \pi_i - \widehat{\boldsymbol{\gamma}}'\boldsymbol{z}_i$ where $\widehat{\boldsymbol{\gamma}}$ is a coefficient estimate from a regression of π_i on $\boldsymbol{z}_i$.

The generated regressor framework includes all of these examples.

The goal is to obtain a distributional approximation for $\widehat{\boldsymbol{\beta}}$ in order to construct standard errors, confidence intervals and conduct tests. Start by substituting equation (11.48) into (11.49). We obtain

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \widehat{\boldsymbol{w}}_i \widehat{\boldsymbol{w}}_i'\right)^{-1} \left(\sum_{i=1}^{n} \widehat{\boldsymbol{w}}_i \left(\boldsymbol{w}_i'\boldsymbol{\beta} + v_i\right)\right).$$

Next, substitute $\boldsymbol{w}_i'\boldsymbol{\beta} = \widehat{\boldsymbol{w}}_i'\boldsymbol{\beta} + (\boldsymbol{w}_i - \widehat{\boldsymbol{w}}_i)'\boldsymbol{\beta}$. We obtain

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\sum_{i=1}^{n} \widehat{\boldsymbol{w}}_i \widehat{\boldsymbol{w}}_i'\right)^{-1} \left(\sum_{i=1}^{n} \widehat{\boldsymbol{w}}_i \left((\boldsymbol{w}_i - \widehat{\boldsymbol{w}}_i)'\boldsymbol{\beta} + v_i\right)\right). \tag{11.50}$$

Effectively, this shows that the distribution of $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ has two random components, one due to the conventional regression component $\widehat{\boldsymbol{w}}_i v_i$, and the second due to the generated regressor $(\boldsymbol{w}_i - \widehat{\boldsymbol{w}}_i)'\boldsymbol{\beta}$. Conventional variance estimators do not address this second component and thus will be biased.

Interestingly, the distribution in (11.50) dramatically simplifies in the special case that the "generated regressor term" $(\boldsymbol{w}_i - \widehat{\boldsymbol{w}}_i)'\boldsymbol{\beta}$ disappears. This occurs when the slope coefficients on the generated regressors are zero. To be specific, partition $\boldsymbol{w}_i = (\boldsymbol{w}_{1i}, \boldsymbol{w}_{2i})$, $\widehat{\boldsymbol{w}}_i = (\boldsymbol{w}_{1i}, \widehat{\boldsymbol{w}}_{2i})$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ so that $\boldsymbol{w}_{1i}$ are the conventional observed regressors and $\widehat{\boldsymbol{w}}_{2i}$ are the generated regressors. Then $(\boldsymbol{w}_i - \widehat{\boldsymbol{w}}_i)'\boldsymbol{\beta} = (\boldsymbol{w}_{2i} - \widehat{\boldsymbol{w}}_{2i})'\boldsymbol{\beta}_2$. Thus if $\boldsymbol{\beta}_2 = \boldsymbol{0}$ this term disappears. In this case (11.50) equals

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \widehat{\boldsymbol{w}}_i \widehat{\boldsymbol{w}}_i'\right)^{-1} \left(\sum_{i=1}^{n} \widehat{\boldsymbol{w}}_i v_i\right).$$

This is a dramatic simplification.

Furthermore, since $\widehat{\boldsymbol{w}}_i = \widehat{\boldsymbol{A}}'\boldsymbol{z}_i$ we can write the estimator as a function of sample moments:

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\widehat{\boldsymbol{A}}'\left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i'\right)\widehat{\boldsymbol{A}}\right)^{-1} \widehat{\boldsymbol{A}}'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \boldsymbol{z}_i v_i\right).$$

If $\widehat{\boldsymbol{A}} \xrightarrow{p} \boldsymbol{A}$ we find from standard manipulations that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}\right)^{-1}\left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'v_i^2\right)\boldsymbol{A}\right)\left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}\right)^{-1}. \tag{11.51}$$

The conventional asymptotic covariance matrix estimator for $\widehat{\boldsymbol{\beta}}$ takes the form

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{w}}_i\widehat{\boldsymbol{w}}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{w}}_i\widehat{\boldsymbol{w}}_i'\widehat{v}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{w}}_i\widehat{\boldsymbol{w}}_i'\right)^{-1} \tag{11.52}$$

where $\widehat{v}_i = y_i - \widehat{\boldsymbol{w}}_i'\widehat{\boldsymbol{\beta}}$. Under the given assumptions, $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$. Thus inference using $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ is asymptotically valid. This is useful when we are interested in tests of $\boldsymbol{\beta}_2 = \boldsymbol{0}$. Often this is of major interest in applications.

To test $\mathbb{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$ we partition $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2\right)$ and construct a conventional Wald statistic

$$W = n\widehat{\boldsymbol{\beta}}_2'\left(\left[\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\right]_{22}\right)^{-1}\widehat{\boldsymbol{\beta}}_2.$$

Theorem 11.22.1 *Take model (11.48) with $\mathbb{E}\left(y_i^4\right) < \infty$, $\mathbb{E}\left\|\boldsymbol{z}_i\right\|^4 < \infty$, $\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A} > 0$, $\widehat{\boldsymbol{A}} \xrightarrow{p} \boldsymbol{A}$ and $\widehat{\boldsymbol{w}}_i = (\boldsymbol{w}_{1i}, \widehat{\boldsymbol{w}}_{2i})$. Under $\mathbb{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$, then as $n \to \infty$,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where $\boldsymbol{V}_{\boldsymbol{\beta}}$ is given in (11.51). For $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ given in (11.52),

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}.$$

Furthermore,

$$W \xrightarrow{d} \chi_q^2$$

where $q = \dim(\boldsymbol{\beta}_2)$. For c satisfying $\alpha = 1 - G_q(c)$

$$\Pr\left(W > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ if $W > c$" has asymptotic size α.

In the special case that $\widehat{\boldsymbol{A}} = \boldsymbol{A}\left(\boldsymbol{X}, \boldsymbol{Z}\right)$ and $v_i|\boldsymbol{x}_i, \boldsymbol{z}_i \sim \mathrm{N}\left(0, \sigma^2\right)$ then there is a finite sample version of the previous result. Let W^0 be the Wald statistic constructed with a homoskedastic variance matrix estimator, and let

$$F = W/q \tag{11.53}$$

be the the F statistic, where $q = \dim(\boldsymbol{\beta}_2)$.

Theorem 11.22.2 *Take model (11.48) with $\widehat{\boldsymbol{A}} = \boldsymbol{A}\left(\boldsymbol{X}, \boldsymbol{Z}\right)$, $v_i|\boldsymbol{x}_i, \boldsymbol{z}_i \sim \mathrm{N}\left(0, \sigma^2\right)$ and $\widehat{\boldsymbol{w}}_i = (\boldsymbol{w}_{1i}, \widehat{\boldsymbol{w}}_{2i})$. Under $\mathbb{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}$, t-statistics have exact $\mathrm{N}(0,1)$ distributions, and the F statistic (11.53) has an exact $F_{q,n-k}$ distribution, where $q = \dim(\boldsymbol{\beta}_2)$ and $k = \dim(\boldsymbol{\beta})$.*

The theory introduced above allows tests of $\mathbb{H}_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ but does not lead to methods to construct standard errors or confidence intervals. For this, we need to work out the distribution without imposing the simplification $\boldsymbol{\beta}_2 = \mathbf{0}$. This often needs to be worked out case-by-case, or by using methods based on the generalized method of moments to be introduced in Chapter 12. However, in some important set of examples it is straightforward to work out the asymptotic distribution.

For the remainder of this section we examine the setting where the estimators $\widehat{\boldsymbol{A}}$ take a least-squares form, so for some $\boldsymbol{X}$ can be written as $\widehat{\boldsymbol{A}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}(\boldsymbol{Z}'\boldsymbol{X})$. Such estimators correspond to the multivariate projection model

$$\begin{aligned} \boldsymbol{x}_i &= \boldsymbol{A}'\boldsymbol{z}_i + \boldsymbol{u}_i \\ \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{u}_i'\right) &= \mathbf{0}. \end{aligned} \tag{11.54}$$

This class of estimators directly includes 2SLS and the expectation model described above. We can write the matrix of generated regressors as $\widehat{\boldsymbol{W}} = \boldsymbol{Z}\widehat{\boldsymbol{A}}$ and then (11.50) as

$$\begin{aligned} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \left(\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\left(\widehat{\boldsymbol{W}}'\left(\left(\boldsymbol{W} - \widehat{\boldsymbol{W}}\right)\boldsymbol{\beta} + \boldsymbol{v}\right)\right) \\ &= \left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\left(-\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{U}\right)\boldsymbol{\beta} + \boldsymbol{v}\right)\right) \\ &= \left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\left(-\boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{v}\right)\right) \\ &= \left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{e}\right) \end{aligned}$$

where

$$e_i = v_i - \boldsymbol{u}_i'\boldsymbol{\beta} = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{11.55}$$

This estimator has the asymptotic distribution

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}\right)^{-1}\left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)\boldsymbol{A}\right)\left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}\right)^{-1}. \tag{11.56}$$

Under conditional homoskedasticity the covariance matrix simplifies to

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}\right)^{-1}\mathbb{E}\left(e_i^2\right).$$

An appropriate estimator of $\boldsymbol{V}_{\boldsymbol{\beta}}$ is

$$\begin{aligned} \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} &= \left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{w}}_i\widehat{\boldsymbol{w}}_i'\widehat{e}_i^2\right)\left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1} \\ \widehat{e}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}. \end{aligned} \tag{11.57}$$

Under the assumption of conditional homoskedasticity this can be simplified as usual.

This appears to be the usual covariance matrix estimator, but it is not, because the least-squares residuals $\widehat{v}_i = y_i - \widehat{\boldsymbol{w}}_i'\widehat{\boldsymbol{\beta}}$ have been replaced with $\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$. This is exactly the substitution made by the 2SLS covariance matrix formula. Indeed, the covariance matrix estimator $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ precisely equals the estimator (11.45).

Theorem 11.22.3 *Take model (11.48) and (11.54) with* $\mathbb{E}\left(y_i^4\right) < \infty$, $\mathbb{E}\left\|\boldsymbol{z}_i\right\|^4 < \infty$, $\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A} > 0$, *and* $\widehat{\boldsymbol{A}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$. *As* $n \to \infty$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where $\boldsymbol{V}_{\boldsymbol{\beta}}$ *is given in (11.56) with* e_i *defined in (11.55). For* $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ *given in (11.57),*

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}.$$

Since the parameter estimates are asymptotically normal and the covariance matrix is consistently estimated, standard errors and test statistics constructed from $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$ are asymptotically valid with conventional interpretations.

We now summarize the results of this section. In general, care needs to be exercised when estimating models with generated regressors. As a general rule, generated regressors and two-step estimation affects sampling distributions and variance matrices. An important simplication occurs for tests that the generated regressors have zero slopes. In this case conventional tests have conventional distributions, both asymptotically and in finite samples. Another important special case occurs when the generated regressors are least-squares fitted values. In this case the asymptotic distribution takes a conventional form, but the conventional residual needs to be replaced by one constructed with the forecasted variable. With this one modification asymptotic inference using the generated regressors is conventional.

11.23 Regression with Expectation Errors

In this section we examine a generated regressor model which includes expectation errors in the regression. This is an important class of generated regressor models, and is relatively straightforward to characterize.

The model is

$$\begin{aligned} y_i &= \boldsymbol{w}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\alpha} + v_i \\ \boldsymbol{w}_i &= \boldsymbol{A}'\boldsymbol{z}_i \\ \boldsymbol{x}_i &= \boldsymbol{w}_i + \boldsymbol{u}_i \\ \mathbb{E}\left(\boldsymbol{z}_i\varepsilon_i\right) &= \boldsymbol{0} \\ \mathbb{E}\left(\boldsymbol{u}_i\varepsilon_i\right) &= \boldsymbol{0} \\ \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{u}_i'\right) &= \boldsymbol{0}. \end{aligned}$$

The observables are $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$. This model states that $\boldsymbol{w}_i$ is the expectation of $\boldsymbol{x}_i$ (or more generally, the projection of $\boldsymbol{x}_i$ on $\boldsymbol{z}_i$) and $\boldsymbol{u}_i$ is its expectation error. The model allows for exogenous regressors as in the standard IV model if they are listed in $\boldsymbol{w}_i$, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$. This model is used, for example, to decompose the effect of expectations from expectation errors. In some cases it is desired to include only the expecation error $\boldsymbol{u}_i$, not the expecation $\boldsymbol{w}_i$. This does not change the results described here.

The model is estimated as follows. First, $\boldsymbol{A}$ is estimated by multivariate least-squares of $\boldsymbol{x}_i$ on $\boldsymbol{z}_i$, $\widehat{\boldsymbol{A}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$, which yields as by-products the fitted values $\widehat{\boldsymbol{W}} = \boldsymbol{Z}\widehat{\boldsymbol{A}}$ and residuals $\widehat{\boldsymbol{U}} = \widehat{\boldsymbol{X}} - \widehat{\boldsymbol{W}}$. Second, the coefficients are estimated by least-squares of y_i on the fitted values $\widehat{\boldsymbol{w}}_i$ and residuals $\widehat{\boldsymbol{u}}_i$

$$y_i = \widehat{\boldsymbol{w}}_i'\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}}_i'\widehat{\boldsymbol{\alpha}} + \widehat{v}_i.$$

We now examine the asymptotic distributions of these estimates.

By the first-step regression $\boldsymbol{Z}'\widehat{\boldsymbol{U}}=\boldsymbol{0}$, $\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{U}}=\boldsymbol{0}$ and $\boldsymbol{W}'\widehat{\boldsymbol{U}}=\boldsymbol{0}$. This means that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ can be computed separately. Notice that

$$\widehat{\boldsymbol{\beta}}=\left(\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\widehat{\boldsymbol{W}}'\boldsymbol{y}$$

and

$$\boldsymbol{y}=\widehat{\boldsymbol{W}}\boldsymbol{\beta}+\boldsymbol{U}\boldsymbol{\alpha}+\left(\boldsymbol{W}-\widehat{\boldsymbol{W}}\right)\boldsymbol{\beta}+\boldsymbol{v}.$$

Substituting, using $\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{U}}=0$ and $\boldsymbol{W}-\widehat{\boldsymbol{W}}=-\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{U}$ we find

$$\begin{aligned}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}&=\left(\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\widehat{\boldsymbol{W}}'\left(\boldsymbol{U}\boldsymbol{\alpha}+\left(\boldsymbol{W}-\widehat{\boldsymbol{W}}\right)\boldsymbol{\beta}+\boldsymbol{v}\right)\\&=\left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\left(\boldsymbol{U}\boldsymbol{\alpha}-\boldsymbol{U}\boldsymbol{\beta}+\boldsymbol{v}\right)\\&=\left(\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{Z}\widehat{\boldsymbol{A}}\right)^{-1}\widehat{\boldsymbol{A}}'\boldsymbol{Z}'\boldsymbol{e}\end{aligned}$$

where

$$e_i=v_i+\boldsymbol{u}_i'\left(\boldsymbol{\alpha}-\boldsymbol{\beta}\right)=y_i-\boldsymbol{x}_i'\boldsymbol{\beta}.$$

We also find

$$\widehat{\boldsymbol{\alpha}}=\left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\widehat{\boldsymbol{U}}'\boldsymbol{y}.$$

Since $\widehat{\boldsymbol{U}}'\boldsymbol{W}=\boldsymbol{0}$, $\boldsymbol{U}-\widehat{\boldsymbol{U}}=\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{U}$ and $\widehat{\boldsymbol{U}}'\boldsymbol{Z}=\boldsymbol{0}$ then

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}&=\left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\widehat{\boldsymbol{U}}'\left(\boldsymbol{W}\boldsymbol{\beta}+\left(\boldsymbol{U}-\widehat{\boldsymbol{U}}\right)\boldsymbol{\alpha}+\boldsymbol{v}\right)\\&=\left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\widehat{\boldsymbol{U}}'\boldsymbol{v}.\end{aligned}$$

Together, we establish the following distributional result.

Theorem 11.23.1 *For the model and estimates described in this section, with* $\mathbb{E}\left(y_i^4\right)<\infty$, $\mathbb{E}\left\|\boldsymbol{z}_i\right\|^4<\infty$, $\mathbb{E}\left\|\boldsymbol{x}_i\right\|^4<\infty$, $\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}>0$, *and* $\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)>0$, *as* $n\to\infty$

$$\sqrt{n}\begin{pmatrix}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \widehat{\boldsymbol{\alpha}}-\boldsymbol{\alpha}\end{pmatrix}\xrightarrow{d}\mathrm{N}\left(\boldsymbol{0},\boldsymbol{V}\right)\tag{11.58}$$

where

$$\boldsymbol{V}=\begin{pmatrix}\boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}&\boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{\alpha}}\\ \boldsymbol{V}_{\boldsymbol{\alpha}\boldsymbol{\beta}}&\boldsymbol{V}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\end{pmatrix}$$

and

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{\beta}\boldsymbol{\beta}}&=\left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}\right)^{-1}\left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)\boldsymbol{A}\right)\left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}\right)^{-1}\\\boldsymbol{V}_{\boldsymbol{\alpha}\boldsymbol{\beta}}&=\left(\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)\right)^{-1}\left(\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{z}_i'e_iv_i\right)\boldsymbol{A}\right)\left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}\right)^{-1}\\\boldsymbol{V}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}&=\left(\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'v_i^2\right)\left(\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)\right)^{-1}.\end{aligned}$$

The asymptotic covariance matrix is estimated by

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta\beta}} = \left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{w}}_i\widehat{\boldsymbol{w}}_i'\widehat{e}_i^2\right)\left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}$$
$$\widehat{\boldsymbol{V}}_{\boldsymbol{\alpha\beta}} = \left(\frac{1}{n}\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{u}}_i\widehat{\boldsymbol{w}}_i'\widehat{e}_i\widehat{v}_i\right)\left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}$$
$$\widehat{\boldsymbol{V}}_{\boldsymbol{\alpha\alpha}} = \left(\frac{1}{n}\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{u}}_i\widehat{\boldsymbol{u}}_i'\widehat{v}_i^2\right)\left(\frac{1}{n}\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}$$

where

$$\begin{aligned}\widehat{\boldsymbol{w}}_i &= \widehat{\boldsymbol{A}}'\boldsymbol{z}_i \\ \widehat{\boldsymbol{u}}_i &= \widehat{\boldsymbol{x}}_i - \widehat{\boldsymbol{w}}_i \\ \widehat{e}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \\ \widehat{v}_i &= y_i - \widehat{\boldsymbol{w}}_i'\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{u}}_i'\widehat{\boldsymbol{\alpha}}.\end{aligned}$$

Under conditional homoskedasticity, specifically

$$\mathbb{E}\left(\begin{pmatrix} e_i^2 & e_iv_i \\ e_iv_i & v_i^2 \end{pmatrix}|\boldsymbol{z}_i\right) = C$$

then $\boldsymbol{V}_{\boldsymbol{\alpha\beta}} = 0$ and the coefficient estimates $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ are asymptotically independent. The variance components also simplify to

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{\beta\beta}} &= \left(\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A}\right)^{-1}\mathbb{E}\left(e_i^2\right) \\ \boldsymbol{V}_{\boldsymbol{\alpha\alpha}} &= \left(\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)\right)^{-1}\mathbb{E}\left(v_i^2\right).\end{aligned}$$

In this case we have the covariance matrix estimators

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta\beta}}^0 = \left(\frac{1}{n}\widehat{\boldsymbol{W}}'\widehat{\boldsymbol{W}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2\right)$$
$$\widehat{\boldsymbol{V}}_{\boldsymbol{\alpha\alpha}}^0 = \left(\frac{1}{n}\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{v}_i^2\right)$$

and $\widehat{\boldsymbol{V}}_{\boldsymbol{\alpha\beta}}^0 = \boldsymbol{0}$.

11.24 Control Function Regression

In this section we present an alternative way of computing the 2SLS estimator by least squares. It is useful in more complicated nonlinear contexts, and also in the linear model to construct tests for endogeneity.

The structural and reduced form equations for the standard IV model are

$$\begin{aligned}y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \boldsymbol{x}_{2i} &= \boldsymbol{\Gamma}_{12}'\boldsymbol{z}_{1i} + \boldsymbol{\Gamma}_{22}'\boldsymbol{z}_{2i} + \boldsymbol{u}_{2i}.\end{aligned}$$

Since the instrumental variable assumption specifies that $\mathbb{E}\left(\boldsymbol{z}_ie_i\right) = \boldsymbol{0}$, $\boldsymbol{x}_{2i}$ is endogenous (correlated with e_i) if and only if $\boldsymbol{u}_{2i}$ and e_i are correlated. We can therefore consider the linear projection of

e_i on $\boldsymbol{u}_{2i}$

$$\begin{aligned} e_i &= \boldsymbol{u}_{2i}'\boldsymbol{\alpha} + \varepsilon_i \\ \boldsymbol{\alpha} &= \left(\mathbb{E}\left(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{u}_{2i}e_i\right) \\ \mathbb{E}\left(\boldsymbol{u}_{2i}\varepsilon_i\right) &= \mathbf{0}. \end{aligned}$$

Substituting this into the structural form equation we find

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \boldsymbol{u}_{2i}'\boldsymbol{\alpha} + \varepsilon_i \qquad (11.59) \\ \mathbb{E}\left(\boldsymbol{x}_{1i}\varepsilon_i\right) &= \mathbf{0} \\ \mathbb{E}\left(\boldsymbol{x}_{2i}\varepsilon_i\right) &= \mathbf{0} \\ \mathbb{E}\left(\boldsymbol{u}_{2i}\varepsilon_i\right) &= \mathbf{0}. \end{aligned}$$

Notice that $\boldsymbol{x}_{2i}$ is uncorrelated with ε_i. This is because $\boldsymbol{x}_{2i}$ is correlated with e_i only through $\boldsymbol{u}_{2i}$, and ε_i is the error after e_i has been projected orthogonal to $\boldsymbol{u}_{2i}$.

If $\boldsymbol{u}_{2i}$ were observed we could then estimate (11.59) by least-squares. While it is not observed, we can estimate $\boldsymbol{u}_{2i}$ by the reduced-form residual

$$\widehat{\boldsymbol{u}}_{2i} = \boldsymbol{x}_{2i} - \widehat{\boldsymbol{\Gamma}}_{12}'\boldsymbol{z}_{1i} - \widehat{\boldsymbol{\Gamma}}_{22}'\boldsymbol{z}_{2i}$$

as defined in (11.23). Then the coefficients $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\alpha})$ can be estimated by least-squares of y_i on $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \widehat{\boldsymbol{u}}_{2i})$. We can write this as

$$y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}}_{2i}'\widehat{\boldsymbol{\alpha}} + \widehat{\varepsilon}_i \qquad (11.60)$$

or in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{U}}_2\widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\varepsilon}}.$$

This turns out to be an alternative algebraic expression for the 2SLS estimator.

Indeed, we now show that $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_{2\text{sls}}$. First, note that the reduced form residual can be written as

$$\widehat{\boldsymbol{U}}_2 = \left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2$$

where $\boldsymbol{P}_{\boldsymbol{Z}}$ is defined in (11.35). By the FWL representation

$$\widehat{\boldsymbol{\beta}} = \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\left(\widetilde{\boldsymbol{X}}'\boldsymbol{y}\right) \qquad (11.61)$$

where $\widetilde{\boldsymbol{X}} = \left[\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2\right]$, with

$$\widetilde{\boldsymbol{X}}_1 = \boldsymbol{X}_1 - \widehat{\boldsymbol{U}}_2\left(\widehat{\boldsymbol{U}}_2'\widehat{\boldsymbol{U}}_2\right)^{-1}\widehat{\boldsymbol{U}}_2'\boldsymbol{X}_1 = \boldsymbol{X}_1$$

(since $\widehat{\boldsymbol{U}}_2'\boldsymbol{X}_1 = 0$) and

$$\begin{aligned} \widetilde{\boldsymbol{X}}_2 &= \boldsymbol{X}_2 - \widehat{\boldsymbol{U}}_2\left(\widehat{\boldsymbol{U}}_2'\widehat{\boldsymbol{U}}_2\right)^{-1}\widehat{\boldsymbol{U}}_2'\boldsymbol{X}_2 \\ &= \boldsymbol{X}_2 - \widehat{\boldsymbol{U}}_2\left(\boldsymbol{X}_2'\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{Z}}\right)\boldsymbol{X}_2 \\ &= \boldsymbol{X}_2 - \widehat{\boldsymbol{U}}_2 \\ &= \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2. \end{aligned}$$

Thus $\widetilde{\boldsymbol{X}} = [\boldsymbol{X}_1, \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2] = \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}$. Substituted into (11.61) we find

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{y}\right) = \widehat{\boldsymbol{\beta}}_{2\text{sls}}$$

which is (11.36) as claimed.

Again, what we have found is that OLS estimation of equation (11.60) yields algebraically the 2SLS estimator $\widehat{\boldsymbol{\beta}}_{2\text{sls}}$.

We now consider the distribution of the control function estimates. It is a generated regression model, and in fact is covered by the model examined in Section 11.23 after a slight reparametrization. Let $\boldsymbol{w}_i = \boldsymbol{\Gamma}'\boldsymbol{z}_i$ and $\boldsymbol{u}_i = \boldsymbol{x}_i - \boldsymbol{\Gamma}'\boldsymbol{z}_i = (\boldsymbol{0}', \boldsymbol{u}_{2i}')'$. Then the main equation (11.59) can be written as

$$y_i = \boldsymbol{w}_i'\boldsymbol{\beta} + \boldsymbol{u}_{2i}'\boldsymbol{\gamma} + \varepsilon_i$$

where $\boldsymbol{\gamma} = \boldsymbol{\alpha} + \boldsymbol{\beta}_2$. This is the model in Section 11.23.

Set $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\beta}}_2$ It follows from (11.58) that as $n \to \infty$ we have the joint distribution

$$\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_2 \\ \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \end{pmatrix} \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V})$$

where

$$\boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{22} & \boldsymbol{V}_{2\boldsymbol{\gamma}} \\ \boldsymbol{V}_{\boldsymbol{\gamma}2} & \boldsymbol{V}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \end{pmatrix}$$

$$\begin{aligned}
\boldsymbol{V}_{22} &= \left[\left(\boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma}\right)^{-1}\left(\boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\boldsymbol{\Gamma}\right)\right)\left(\boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma}\right)^{-1}\right]_{22} \\
\boldsymbol{V}_{\boldsymbol{\gamma}2} &= \left[\left(\mathbb{E}\left(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}'\right)\right)^{-1}\left(\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{z}_i'e_i\varepsilon_i\right)\boldsymbol{\Gamma}\right)\left(\boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma}\right)^{-1}\right]_{\cdot 2} \\
\boldsymbol{V}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} &= \left(\mathbb{E}\left(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}'\varepsilon_i^2\right)\left(\mathbb{E}\left(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}'\right)\right)^{-1} \\
e_i &= y_i - \boldsymbol{x}_i'\boldsymbol{\beta}.
\end{aligned}$$

The asymptotic distribution of $\widehat{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\alpha}} - \widehat{\boldsymbol{\beta}}_2$ can then be deduced.

Theorem 11.24.1 *If* $\mathbb{E}\left(y_i^4\right) < \infty$, $\mathbb{E}\left\|\boldsymbol{z}_i\right\|^4 < \infty$, $\mathbb{E}\left\|\boldsymbol{x}_i\right\|^4 < \infty$, $\boldsymbol{A}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{A} > 0$, *and* $\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right) > 0$, *as* $n \to \infty$

$$\sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\alpha}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\alpha}} = \boldsymbol{V}_{22} + \boldsymbol{V}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} - \boldsymbol{V}_{2\boldsymbol{\gamma}} - \boldsymbol{V}_{\boldsymbol{\gamma}2}.$$

$$\sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\alpha}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\alpha}} = \boldsymbol{V}_{22} + \boldsymbol{V}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} - \boldsymbol{V}_{2\boldsymbol{\gamma}} - \boldsymbol{V}_{\boldsymbol{\gamma}2}.$$

Under conditional homoskedasticity we have the important simplifications

$$\begin{aligned}
\boldsymbol{V}_{22} &= \left[\left(\boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma}\right)^{-1}\right]_{22}\mathbb{E}\left(e_i^2\right) \\
\boldsymbol{V}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} &= \left(\mathbb{E}\left(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}'\right)\right)^{-1}\mathbb{E}\left(\varepsilon_i^2\right) \\
\boldsymbol{V}_{\boldsymbol{\gamma}2} &= \boldsymbol{0} \\
\boldsymbol{V}_{\boldsymbol{\alpha}} &= \boldsymbol{V}_{22} + \boldsymbol{V}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}.
\end{aligned}$$

An estimator for $\boldsymbol{V}_{\boldsymbol{\alpha}}$ in the general case is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\alpha}} = \widehat{\boldsymbol{V}}_{22} + \widehat{\boldsymbol{V}}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} - \widehat{\boldsymbol{V}}_{2\boldsymbol{\gamma}} - \widehat{\boldsymbol{V}}_{\boldsymbol{\gamma}2} \tag{11.62}$$

where

$$\widehat{\boldsymbol{V}}_{22} = \left[\frac{1}{n}\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}_i'\widehat{e}_i^2\right)\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\right]_{22}$$

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\gamma}2} = \left[\frac{1}{n}\left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\sum_{i=1}^{n}\widehat{\boldsymbol{u}}_i\widehat{\boldsymbol{w}}_i'\widehat{e}_i\widehat{\varepsilon}_i\right)\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\right]_{\cdot 2}$$

$$\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$$

$$\widehat{\varepsilon}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{u}}_{2i}'\widehat{\boldsymbol{\alpha}}.$$

Under the assumption of conditional homoskedasticity we have the estimator

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\alpha}}^0 = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}\boldsymbol{\beta}}^0 + \widehat{\boldsymbol{V}}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^0$$

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \left[\left(\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\right]_{22}\left(\sum_{i=1}^{n}\widehat{e}_i^2\right)$$

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = \left(\widehat{\boldsymbol{U}}'\widehat{\boldsymbol{U}}\right)^{-1}\left(\sum_{i=1}^{n}\widehat{\varepsilon}_i^2\right).$$

11.25 Endogeneity Tests

The 2SLS estimator allows the regressor $\boldsymbol{x}_{2i}$ to be endogenous, meaning that $\boldsymbol{x}_{2i}$ is correlated with the structural error e_i. If this correlation is zero, then $\boldsymbol{x}_{2i}$ is exogenous and the structural equation can be estimated by least-squares. This is a testable restriction. Effectively, the null hypothesis is

$$\mathbb{H}_0 : \mathbb{E}(\boldsymbol{x}_{2i}e_i) = \boldsymbol{0}$$

with the alternative

$$\mathbb{H}_1 : \mathbb{E}(\boldsymbol{x}_{2i}e_i) \neq \boldsymbol{0}.$$

The maintained hypothesis is $\mathbb{E}(\boldsymbol{z}_i e_i) = \boldsymbol{0}$. Since $\boldsymbol{x}_{1i}$ is a component of $\boldsymbol{z}_i$, this implies $\mathbb{E}(\boldsymbol{x}_{1i}e_i) = \boldsymbol{0}$. Consequently we could alternatively write the null as $\mathbb{H}_0 : \mathbb{E}(\boldsymbol{x}_i e_i) = \boldsymbol{0}$ (and some authors do so).

Recall the control function regression (11.59)

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \boldsymbol{u}_{2i}'\boldsymbol{\alpha} + \varepsilon_i$$

$$\boldsymbol{\alpha} = \left(\mathbb{E}\left(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{u}_{2i}e_i\right).$$

Notice that $\mathbb{E}(\boldsymbol{x}_{2i}e_i) = \boldsymbol{0}$ if and only if $\mathbb{E}\left(\boldsymbol{u}_{2i}e_i\right) = \boldsymbol{0}$, so the hypothesis can be restated as $\mathbb{H}_0 : \boldsymbol{\alpha} = \boldsymbol{0}$ against $\mathbb{H}_1 : \boldsymbol{\alpha} \neq \boldsymbol{0}$. Thus a natural test is based on the Wald statistic W for $\boldsymbol{\alpha} = \boldsymbol{0}$ in the control function regression (11.24). Under Theorem 11.22.1 and Theorem 11.22.2, under $\mathbb{H}_0$, W is asymptotically chi-square with k_2 degrees of freedom. In addition, under the normal regression assumptions the F statistic has an exact $F(k_2, n - k_1 - 2k_2)$ distribution. We accept the null hypothesis that $\boldsymbol{x}_{2i}$ is exogenous if W (or F) is smaller than the critical value, and reject in favor of the hypothesis that $\boldsymbol{x}_{2i}$ is endogenous if the statistic is larger than the critical value.

Specifically, estimate the reduced form by least squares

$$\boldsymbol{x}_{2i} = \widehat{\boldsymbol{\Gamma}}_{12}'\boldsymbol{z}_{1i} + \widehat{\boldsymbol{\Gamma}}_{22}'\boldsymbol{z}_{2i} + \widehat{\boldsymbol{u}}_{2i}$$

to obtain the residuals. Then estimate the control function by least squares

$$y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}}_{2i}'\widehat{\boldsymbol{\alpha}} + \widehat{\varepsilon}_i. \tag{11.63}$$

Let W, W^0 and $F = W^0/k_2$ denote the Wald statistic, homoskedastic Wald statistic, and F statistic for $\boldsymbol{\alpha} = \boldsymbol{0}$.

Theorem 11.25.1 *Under* $\mathbb{H}_0$, $W \xrightarrow{d} \chi^2_{k_2}$. *Let* $c_{1-\alpha}$ *solve* $\Pr\left(\chi^2_{k_2} \leq c_{1-\alpha}\right) = 1-\alpha$. *The test "Reject* $\mathbb{H}_0$ *if* $W > c_{1-\alpha}$*" has asymptotic size* α.

Theorem 11.25.2 *Suppose* $e_i|\boldsymbol{x}_i, \boldsymbol{z}_i \sim \mathrm{N}\left(0, \sigma^2\right)$. *Under* $\mathbb{H}_0$, $F \sim F(k_2, n-k_1-2k_2)$. *Let* $c_{1-\alpha}$ *solve* $\Pr\left(F(k_2, n-k_1-2k_2) \leq c_{1-\alpha}\right) = 1-\alpha$. *The test "Reject* $\mathbb{H}_0$ *if* $F > c_{1-\alpha}$*" has exact size* α.

Since in general we do not want to impose homoskedasticity, these results suggest that the most appropriate test is the Wald statistic constructed with the robust heteroskedastic covariance matrix. This can be computed in Stata using the command `estat endogenous` after `ivregress` when the latter uses a robust covariance option. Stata reports the Wald statistic in F form (and thus uses the F distribution to calculate the p-value) as "Robust regression F". Using the F rather than the χ^2 distribution is not formally justified but is a reasonable finite sample adjustment. If the command `estat endogenous` is applied after `ivregress` without a robust covariance option, Stata reports the F statistic as "Wu-Hausman F".

There is an alternative (and traditional) way to derive a test for endogeneity. Under $\mathbb{H}_0$, both OLS and 2SLS are consistent estimators. But under $\mathbb{H}_1$, they converge to different values. Thus the difference between the OLS and 2SLS estimators is a valid test statistic for endogeneity. It also measures what we often care most about – the impact of endogeneity on the parameter estimates. This literature was developed under the assumption of conditional homoskedasticity (and it is important for these results) so we assume this condition for the development of the statistics.

Let $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2\right)$ be the OLS estimator and let $\widetilde{\boldsymbol{\beta}} = \left(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_2\right)$ be the 2SLS estimator. Under $\mathbb{H}_0$ (and homoskedasticity) the OLS estimator is Gauss-Markov efficient, so by the Hausman equality

$$\begin{aligned}\operatorname{var}\left(\widehat{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_2\right) &= \operatorname{var}\left(\widetilde{\boldsymbol{\beta}}_2\right) - \operatorname{var}\left(\widehat{\boldsymbol{\beta}}_2\right) \\ &= \left(\left(\boldsymbol{X}_2'\left(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1\right)\boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\right)\sigma^2\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$, $\boldsymbol{P}_1 = \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$, and $\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{P}_1$. Thus a valid test statistic for $\mathbb{H}_0$ is

$$T = \frac{\left(\widehat{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_2\right)'\left(\left(\boldsymbol{X}_2'\left(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1\right)\boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\right)^{-1}\left(\widehat{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_2\right)}{\widehat{\sigma}^2} \tag{11.64}$$

for some estimate $\widehat{\sigma}^2$ of σ^2. Durbin (1954) first proposed T as a test for endogeneity in the context of IV estimation, setting $\widehat{\sigma}^2$ to be the least-squares estimate of σ^2. Wu (1973) proposed T as a test for endogeneity in the context of 2SLS estimation, considering a set of possible estimates $\widehat{\sigma}^2$, including the regression estimate from (11.63). Hausman (1978) proposed a version of T based on the full contrast $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$, and observed that it equals the regression Wald statistic W^0 described earlier. In fact, when $\widehat{\sigma}^2$ is the regression estimate from (11.63), the statistic (11.64) algebraically equals both W^0 and the version of (11.64) based on the full contrast $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$. We show these equalities below. Thus these three approaches yield exactly the same statistic except for possible differences regarding the choice of $\widehat{\sigma}^2$. Since the regression F test described earlier has an exact F distribution in the normal sampling model, and thus can exactly control test size, this is the

preferred version of the test. The general class of tests are called **Durbin-Wu-Hausman** tests, **Wu-Hausman** tests, or **Hausman** tests, depending on the author.

When $k_2 = 1$ (there is one right-hand-side endogenous variable) which is quite common in applications, the endogeneity test can be equivalently expressed at the t-statistic for $\widehat{\alpha}$ in the estimated control function. Thus it is sufficient to estimate the control function regression and check the t-statistic for $\widehat{\alpha}$. If $|\widehat{\alpha}| > 2$ then we can reject the hypothesis that $\boldsymbol{x}_{2i}$ is exogenous for $\boldsymbol{\beta}$.

We illustrate using the Card proximity example using the two instruments *public* and *private*. We first estimate the reduced form for *education*, obtain the residual, and then estimate the control function regression. The residual has a coefficient -0.088 with a standard error of 0.037 and a t-statistic of 2.4. Since the latter exceeds the 5% crtical value (its p-value is 0.017) we reject exogeneity. This means that the 2SLS estimates are statistically different from the least-squares estimates of the structural equation and supports our decision to treat education as an endogenous variable. (Alternatively, the F statistic is $2.4^2 = 5.7$ with the same p-value).

We now show the equality of the various statistics.

We first show that the statistic (11.64) is not altered if based on the full contrast $\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}$. Indeed, $\widehat{\boldsymbol{\beta}}_1 - \widetilde{\boldsymbol{\beta}}_1$ is a linear function of $\widehat{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_2$, so there is no extra information in the full contrast. To see this, observe that given $\widehat{\boldsymbol{\beta}}_2$, we can solve by least-squares to find

$$\widehat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\left(\boldsymbol{y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right)\right)$$

and similarly

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_1 &= \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\left(\boldsymbol{y} - \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2\widetilde{\boldsymbol{\beta}}\right)\right) \\ &= \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\left(\boldsymbol{y} - \boldsymbol{X}_2\widetilde{\boldsymbol{\beta}}\right)\right)\end{aligned}$$

the second equality since $\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_1 = \boldsymbol{X}_1$. Thus

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_1 - \widetilde{\boldsymbol{\beta}}_1 &= \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\left(\boldsymbol{y} - \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2\right) - \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\left(\boldsymbol{y} - \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2\widetilde{\boldsymbol{\beta}}\right) \\ &= \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\widetilde{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_2\right)\end{aligned}$$

as claimed.

We next show that T in (11.64) equals the homoskedastic Wald statistic W^0 for $\widehat{\boldsymbol{\alpha}}$ from the regression (11.63). Consider the latter regression. Since $\boldsymbol{X}_2$ is contained in $\boldsymbol{X}$, the coefficient estimate $\widehat{\boldsymbol{\alpha}}$ is invariant to replacing $\widehat{\boldsymbol{U}}_2 = \boldsymbol{X}_2 - \widehat{\boldsymbol{X}}_2$ with $-\widehat{\boldsymbol{X}}_2 = -\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2$. By the FWL representation, setting $\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= -\left(\widehat{\boldsymbol{X}}_2'\boldsymbol{M}_{\boldsymbol{X}}\widehat{\boldsymbol{X}}_2\right)^{-1}\widehat{\boldsymbol{X}}_2'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} \\ &= -\left(\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}.\end{aligned} \tag{11.65}$$

It follows that

$$W^0 = \frac{\boldsymbol{y}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}}{\widehat{\sigma}^2}.$$

Our goal is to show that $T = W^0$. Define $\widetilde{\boldsymbol{X}}_2 = \left(\boldsymbol{I}_n - \boldsymbol{P}_1\right)\boldsymbol{X}_2$ so $\widehat{\boldsymbol{\beta}}_2 = \left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\widetilde{\boldsymbol{X}}_2'\boldsymbol{y}$. Then

defining using $(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)(\boldsymbol{I}_n - \boldsymbol{P}_1) = (\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)$ and defining $\boldsymbol{Q} = \widetilde{\boldsymbol{X}}_2 \left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1} \widetilde{\boldsymbol{X}}_2'$

$$\begin{aligned}
\boldsymbol{\Delta} &\overset{def}{=} \left(\boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)\boldsymbol{X}_2\right)\left(\widetilde{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_2\right) \\
&= \boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)\boldsymbol{y} - \left(\boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)\boldsymbol{X}_2\right)\left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\widetilde{\boldsymbol{X}}_2'\boldsymbol{y} \\
&= \boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)(\boldsymbol{I}_n - \boldsymbol{Q})\boldsymbol{y} \\
&= \boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1 - \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Q})\boldsymbol{y} \\
&= \boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}(\boldsymbol{I}_n - \boldsymbol{P}_1 - \boldsymbol{Q})\boldsymbol{y} \\
&= \boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}.
\end{aligned}$$

The third-to-last equality is $\boldsymbol{P}_1\boldsymbol{Q} = \boldsymbol{0}$ and the final uses $\boldsymbol{M}_{\boldsymbol{X}} = \boldsymbol{I}_n - \boldsymbol{P}_1 - \boldsymbol{Q}$. We also calculate that

$$\begin{aligned}
\boldsymbol{Q}^* &\overset{def}{=} \left(\boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)\boldsymbol{X}_2\right)\left(\left(\boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)\boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\right) \\
&\cdot \left(\boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)\boldsymbol{X}_2\right) \\
&= \boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1 - (\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1)\boldsymbol{Q}(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1))\boldsymbol{X}_2 \\
&= \boldsymbol{X}_2'(\boldsymbol{P}_{\boldsymbol{Z}} - \boldsymbol{P}_1 - \boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{Q}\boldsymbol{P}_{\boldsymbol{Z}})\boldsymbol{X}_2 \\
&= \boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2.
\end{aligned}$$

Thus

$$\begin{aligned}
T &= \frac{\boldsymbol{\Delta}'\boldsymbol{Q}^{*-1}\boldsymbol{\Delta}}{\widehat{\sigma}^2} \\
&= \frac{\boldsymbol{y}'\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}}{\widehat{\sigma}^2} \\
&= W^0
\end{aligned}$$

as claimed.

11.26 Subset Endogeneity Tests

In some cases we may only wish to test the endogeneity of a subset of the variables. In the Card proximity example, we may wish test the exogeneity of *education* separately from *experience* and its square. To execute a subset endogeneity test it is useful to partition the regressors into three groups, so that the structural model is

$$\begin{aligned}
y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \boldsymbol{x}_{3i}'\boldsymbol{\beta}_3 + e_i \\
\mathbb{E}(\boldsymbol{z}_i e_i) &= \boldsymbol{0}.
\end{aligned}$$

As before, the instrument vector $\boldsymbol{z}_i$ includes $\boldsymbol{x}_{1i}$. The variables $\boldsymbol{x}_{3i}$ is treated as endogenous, and $\boldsymbol{x}_{2i}$ is treated as potentially endogenous. The hypothesis to test is that $\boldsymbol{x}_{2i}$ is exogenous, or

$$\mathbb{H}_0 : \mathbb{E}(\boldsymbol{x}_{2i}e_i) = \boldsymbol{0}$$

against

$$\mathbb{H}_1 : \mathbb{E}(\boldsymbol{x}_{2i}e_i) \neq \boldsymbol{0}.$$

Under homoskedasticity, a straightfoward test can be constructed by the Durbin-Wu-Hausman principle. Under $\mathbb{H}_0$, the appropriate estimator is 2SLS using the instruments $(\boldsymbol{z}_i, \boldsymbol{x}_{2i})$. Let this estimator of $\boldsymbol{\beta}_2$ be denoted $\widehat{\boldsymbol{\beta}}_2$. Under $\mathbb{H}_1$, the appropriate estimator is 2SLS using the smaller

instrument set $\boldsymbol{z}_i$. Let this estimator of $\boldsymbol{\beta}_2$ be denoted $\widetilde{\boldsymbol{\beta}}_2$. A Durbin-Wu-Hausman-type test of $\mathbb{H}_0$ against $\mathbb{H}_1$ is

$$T = \left(\widehat{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_2\right)' \left(\widehat{\operatorname{var}}\left(\widetilde{\boldsymbol{\beta}}_2\right) - \widehat{\operatorname{var}}\left(\widehat{\boldsymbol{\beta}}_2\right)\right)^{-1} \left(\widehat{\boldsymbol{\beta}}_2 - \widetilde{\boldsymbol{\beta}}_2\right).$$

The asymptotic distribution under $\mathbb{H}_0$ is $\chi^2_{k_2}$ where $k_2 = \dim(\boldsymbol{x}_{2i})$, so we reject the hypothesis that the variables $\boldsymbol{x}_{2i}$ are exogenous if T exceeds an upper critical value from the $\chi^2_{k_2}$ distribution.

Instead of using the Wald statistic, one could use the F version of the test by dividing by k_2 and using the F distribution for critical values. There is no finite sample justification for this modification, however, since $\boldsymbol{x}_{3i}$ is endogenous under the null hypothesis.

In Stata, the command `estat endogenous` (adding the variable name to specify which variable to test for exogeneity) after `ivregress` without a robust covariance option reports the F version of this statistic as "Wu-Hausman F". For example, in the Card proximity example using the four instruments *public*, *private*, *age* and age^2, if we estimate the equation by 2SLS with a non-robust covariance matrix, and then compute the endogeneity test for education, we find $F = 272$ with a p-value of 0.0000, but if we compute the test for experience and its square we find $F = 2.98$ with a p-value of 0.051. In this equation, education is clearly endogenous but the experience variables are unclear.

A heteroskedasticity or cluster-robust test cannot be constructed easily by the Durbin-Wu-Hausman approach, since the covariance matrix does not take a simple form. Instead, we can use the regression approach if we account for the generated regressor problem.The ideal control function regression takes the form

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}_{2i}'\boldsymbol{\alpha}_2 + \boldsymbol{u}_{3i}'\boldsymbol{\alpha}_3 + \varepsilon_i$$

where $\boldsymbol{u}_{2i}$ and $\boldsymbol{u}_{3i}$ are the reduced-form errors from the projections of $\boldsymbol{x}_{2i}$ and $\boldsymbol{x}_{3i}$ on the instruments $\boldsymbol{z}_i$. The coefficients $\boldsymbol{\alpha}_2$ and $\boldsymbol{\alpha}_3$ solve the equations

$$\begin{pmatrix} \mathbb{E}(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}') & \mathbb{E}(\boldsymbol{u}_{2i}\boldsymbol{u}_{3i}') \\ \mathbb{E}(\boldsymbol{u}_{3i}\boldsymbol{u}_{2i}') & \mathbb{E}(\boldsymbol{u}_{3i}\boldsymbol{u}_{3i}') \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_3 \end{pmatrix} = \begin{pmatrix} \mathbb{E}(\boldsymbol{u}_{2i}e_i) \\ \mathbb{E}(\boldsymbol{u}_{3i}e_i) \end{pmatrix}.$$

The null hypothesis $\mathbb{E}(\boldsymbol{x}_{2i}e_i) = \boldsymbol{0}$ is equivalent to $\mathbb{E}(\boldsymbol{u}_{2i}e_i) = \boldsymbol{0}$. This implies

$$\boldsymbol{\Psi}' \begin{pmatrix} \boldsymbol{\alpha}_2 \\ \boldsymbol{\alpha}_3 \end{pmatrix} = \boldsymbol{0} \tag{11.66}$$

where

$$\boldsymbol{\Psi} = \begin{pmatrix} \mathbb{E}(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}') \\ \mathbb{E}(\boldsymbol{u}_{3i}\boldsymbol{u}_{2i}') \end{pmatrix}.$$

This suggests that an appropriate regression-based test of $\mathbb{H}_0$ versus $\mathbb{H}_1$ is to construct a Wald statistic for the restriction (11.66) in the control function regression

$$y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}}_{2i}'\widehat{\boldsymbol{\alpha}}_2 + \widehat{\boldsymbol{u}}_{3i}'\widehat{\boldsymbol{\alpha}}_3 + \widehat{\varepsilon}_i \tag{11.67}$$

where $\widehat{\boldsymbol{u}}_{2i}$ and $\widehat{\boldsymbol{u}}_{3i}$ are the least-squares residuals from the regressions of $\boldsymbol{x}_{2i}$ and $\boldsymbol{x}_{3i}$ on the instruments $\boldsymbol{z}_i$, respectively, and $\boldsymbol{\Psi}$ is estimated by

$$\widehat{\boldsymbol{\Psi}} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{u}}_{2i}\widehat{\boldsymbol{u}}_{2i}') \\ \frac{1}{n}\sum_{i=1}^n \widehat{\boldsymbol{u}}_{3i}\widehat{\boldsymbol{u}}_{2i}' \end{pmatrix}.$$

A complication is that the regression (11.67) has generated regressors which have non-zero coefficients under $\mathbb{H}_0$. The solution is to use the control-function-robust covariance matrix estimator (11.62) for $(\widehat{\boldsymbol{\alpha}}_2, \widehat{\boldsymbol{\alpha}}_3)$. This yields a valid Wald statistic for $\mathbb{H}_0$ versus $\mathbb{H}_1$. The asymptotic distribution of the statistic under $\mathbb{H}_0$ is $\chi^2_{k_2}$ where $k_2 = \dim(\boldsymbol{x}_{2i})$, so the null hypothesis that $\boldsymbol{x}_{2i}$ is exogenous is rejected if the Wald statistic exceeds the upper critical value from the $\chi^2_{k_2}$ distribution.

Heteroskedasticity-robust and cluster-robust subset endogeneity tests are not currently implemented in Stata.

11.27 OverIdentification Tests

When $\ell > k$ the model is **overidentified** meaning that there are more moments than free parameters. This is a restriction and is testable. Such tests are callled **overidentification tests**.

The instrumental variables model specifies that

$$\mathbb{E}\left(\boldsymbol{z}_i e_i\right) = \mathbf{0}.$$

Equivalently, since $e_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, this is the same as

$$\mathbb{E}\left(\boldsymbol{z}_i y_i\right) - \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right)\boldsymbol{\beta} = \mathbf{0}.$$

This is an $\ell \times 1$ vector of restrictions on the moment matrices $\mathbb{E}\left(\boldsymbol{z}_i y_i\right)$ and $\mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right)$. Yet since $\boldsymbol{\beta}$ is of dimension k which is less than ℓ, it is not certain if indeed such a $\boldsymbol{\beta}$ exists.

To make things a bit more concrete, suppose there is a single endogenous regressor x_{2i}, no x_{1i}, and two instruments z_{1i} and z_{2i}. Then the model specifies that

$$\mathbb{E}(z_{1i} y_i) = \mathbb{E}(z_{1i} x_{2i})\beta$$

and

$$\mathbb{E}(z_{2i} y_i) = \mathbb{E}(z_{2i} x_{2i})\beta.$$

Thus β solves both equations. This is rather special.

Another way of thinking about this is that in this context we could solve for β using either one equation or the other. In terms of estimation, this is equivalent to estimating by IV using just the instrument z_1 or instead just using the instrument z_2. These two estimators (in finite samples) will be different. But if the overidentification hypothesis is correct, both are estimating the same parameter, and both are consistent for β (if the instruments are relevant). In contrast, if the overidentification hypothesis is false, then the two estimators will converge to different probability limits and it is unclear if either probability limit is interesting.

For example, take the 2SLS estimates in the fourth column of Table 11.1, which use *public* and *private* as instruments for *education*. Suppose we instead estimate by IV, using just *public* as an instrument, and then repeat using *private*. The IV coefficient for *education* in the first case is 0.17, and in the second case 0.27. These appear to be quite different. However, the second estimate has quite a large standard error (0.17) so perhaps the difference is sampling variation. An overidentification test addresses this question formally.

For a general overidentification test, the null and alternative hypotheses are

$$\begin{aligned}\mathbb{H}_0 &: \mathbb{E}(\boldsymbol{z}_i e_i) = \mathbf{0} \\ \mathbb{H}_1 &: \mathbb{E}(\boldsymbol{z}_i e_i) \neq \mathbf{0}.\end{aligned}$$

We will also add the conditional homoskedasticity assumption

$$\mathbb{E}(e_i^2 | \boldsymbol{z}_i) = \sigma^2. \tag{11.68}$$

To avoid imposing (11.68), it is best to take a GMM approach, which we defer until Chapter 12.

To implement a test of $\mathbb{H}_0$, consider a linear regression of the error e_i on the instruments $\boldsymbol{z}_i$

$$e_i = \boldsymbol{z}_i'\boldsymbol{\alpha} + \varepsilon_i \tag{11.69}$$

with

$$\boldsymbol{\alpha} = \left(\mathbb{E}(\boldsymbol{z}_i \boldsymbol{z}_i')\right)^{-1} \mathbb{E}(\boldsymbol{z}_i e_i).$$

We can rewrite $\mathbb{H}_0$ as $\boldsymbol{\alpha} = \mathbf{0}$. While e_i is not observed we can replace it with the 2SLS residual $\widehat{e}_i$, and estimate $\boldsymbol{\alpha}$ by least-squares regression

$$\widehat{\boldsymbol{\alpha}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\widehat{\boldsymbol{e}}.$$

Sargan (1958) proposed testing $\mathbb{H}_0$ via a score test, which takes the form

$$S = \widehat{\boldsymbol{\alpha}}' \left(\widehat{\operatorname{var}}\left(\widehat{\boldsymbol{\alpha}}\right)\right)^{-} \widehat{\boldsymbol{\alpha}} = \frac{\widehat{\boldsymbol{e}}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}}}{\widehat{\sigma}^2}. \tag{11.70}$$

where $\widehat{\sigma}^2 = \frac{1}{n}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$. Basmann (1960) independently proposed a Wald statistic for $\mathbb{H}_0$, which is S with $\widehat{\sigma}^2$ replaced with $\widetilde{\sigma}^2 = n^{-1}\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}$ where $\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{e}} - \boldsymbol{Z}\widehat{\boldsymbol{\alpha}}$. By the equivalence of homoskedastic score and Wald tests (see Section 9.16), Basmann's statistic is a monotonic function of Sargan's statistic and hence they yield equivalent tests. Sargan's version is more typically reported.

The Sargan test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $S > c$ for some critical value c. An asymptotic test sets c as the $1-\alpha$ quantile of the $\chi^2_{\ell-k}$ distribution. This is justified by the asymptotic null distribution of S which we now derive.

Theorem 11.27.1 *Under Assumption 11.14.1 and* $\mathbb{E}(e_i^2|\boldsymbol{z}_i) = \sigma^2$*, then as* $n \to \infty$

$$S \xrightarrow{d} \chi^2_{\ell-k}.$$

For c *satisfying* $\alpha = 1 - G_{\ell-k}(c)$,

$$\Pr\left(S > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ *if* $S > c$*" has asymptotic size* α.

We prove Theorem 11.27.1 below.

The Sargan statistic S is an asymptotic test of the overidentifying restrictions under the assumption of conditional homoskedasticity. It has some limitations. First, it is an asymptotic test, and does not have a finite sample (e.g. F) counterpart. Simulation evidence suggests that the test can be oversized (reject too frequently) in small and moderate sample sizes. Consequently, p-values should be interpreted cautiously. Second, the assumption of conditional homoskedasticity is unrealistic in applications. The best way to generalize the Sargan statistic to allow heteroskedasticity is to use the GMM overidentification statistic – which we will examine in Chapter 12. For 2SLS, Wooldrige (1995) suggested a robust score test, but Baum, Schaffer and Stillman (2003) point out that it is numerically equivalent to the GMM overidentification statistic. Hence the bottom line appears to be that to allow heteroskedasticity or clustering, it is best to use a GMM approach.

In overidentified applications, it is always prudent to report an overidentification test. If the test is insignificant it means that the overidentifying restrictions are not rejected, supporting the estimated model. If the overidentifying test statistic is highly significant (if the p-value is very small) this is evidence that the overidentifying restrictions are violated. In this case we should be concerned that the model is misspecified and interpreting the parameter estimates should be done cautiously.

When reporting the results of an overidentification test, it seems reasonable to focus on very small sigificance levels, such as 1%. This means that we should only treat a model as "rejected" if the Sargan p-value is very small, e.g. less than 0.01. The reason to focus on very small significance levels is because it is very difficult to interpret the result "The model is rejected". Stepping back a bit, it does not seem credible that any overidentified model is literally true, rather what seems potentially credible is that an overidentified model is a reasonable approximation. A test is asking the question "Is there evidence that a model is not true" when we really want to know the answer to "Is there evidence that the model is a poor approximation". Consequently it seems reasonable to require strong evidence to lead to the conclusion "Let's reject this model". The recommendation is that mild rejections (p-values between 1% and 5%) should be viewed as mildly worrisome, but

not critical evidence against a model. The results of an overidentification test should be integrated with other information before making a strong decision.

We illustrate the methods with the Card college proximity example. We have estimated two overidentified models by 2SLS, in columns 4 & 5 of Table 11.1. In each case, the number of overidentifying restrictions is 1. We report the Sargan statistic and its asymptotic p-value (calculated using the χ^2_1 distribution) in the table. Both p-values (0.36 and 0.52) are far from significant, indicating that there is no evidence that the models are misspecified.

We now prove Theorem 11.27.1. The statistic S is invariant to rotations of $\boldsymbol{Z}$ (replacing $\boldsymbol{Z}$ with $\boldsymbol{Z}\boldsymbol{C}$) so without loss of generality we assume $\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right) = \boldsymbol{I}_\ell$. As $n \to \infty$, $n^{-1/2}\boldsymbol{Z}'\boldsymbol{e} \xrightarrow{d} \sigma \mathrm{Z}$ where $\mathrm{Z} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_\ell\right)$. Also $\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z} \xrightarrow{p} \boldsymbol{I}_\ell$ and $\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{Q}$, say. Then

$$n^{-1/2}\boldsymbol{Z}'\widehat{\boldsymbol{e}} = \left(\boldsymbol{I}_\ell - \left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{Z}}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right) n^{-1/2}\boldsymbol{Z}'\boldsymbol{e}$$

$$\xrightarrow{d} \sigma\left(\boldsymbol{I}_\ell - \boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\right)\mathrm{Z}.$$

Since $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$ it follows that

$$S \xrightarrow{d} \mathrm{Z}'\left(\boldsymbol{I}_\ell - \boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\right)\mathrm{Z} \sim \chi^2_{\ell-k}.$$

The distribution is $\chi^2_{\ell-k}$ since $\boldsymbol{I}_\ell - \boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'$ is idempotent with rank $\ell - k$.

The Sargan statistic test can be implemented in Stata using the command `estat overid` after `ivregress 2sls` or `ivregres liml` if a standard (non-robust) covariance matrix has been specified (that is, without the '`,r`' option), or by the command `estat overid, forcenonrobust` otherwise.

11.28 Subset OverIdentification Tests

Tests of $\mathbb{H}_0 : \mathbb{E}(\boldsymbol{z}_i e_i) = \boldsymbol{0}$ are typically interpreted as tests of model specification. The alternative $\mathbb{H}_1 : \mathbb{E}(\boldsymbol{z}_i e_i) \neq \boldsymbol{0}$ means that at least one element of $\boldsymbol{z}_i$ is correlated with the error e_i and is thus an invalid instrumental variable. In some cases it may be reasonable to test only a subset of the moment conditions.

As in the previous section we restrict attention to the homoskedasticity case $\mathbb{E}(e_i^2|\boldsymbol{z}_i) = \sigma^2$.

Partition $\boldsymbol{z}_i = (\boldsymbol{z}_{ai}, \boldsymbol{z}_{bi})$ with dimensions ℓ_a and ℓ_b, respectively, where $\boldsymbol{z}_{ai}$ contains the instruments which are believed to be uncorrelated with e_i, and $\boldsymbol{z}_{bi}$ contains the instruments which may be correlated with e_i. It is necessary to select this partition so that $\ell_a > k$, or equivalently $\ell_b < \ell - k$. This means that the model with just the instruments $\boldsymbol{z}_{ai}$ is over-identified, or that ℓ_b is smaller than the number of overidentifying restrictions. (If $\ell_a = k$ then the tests described here exist but reduce to the Sargan test so are not interesting.) Hence the tests require that $\ell - k > 1$, that the number of overidentifying restrictions exceeds one.

Given this partition, the maintained hypothesis is that $\mathbb{E}(\boldsymbol{z}_{ai} e_i) = \boldsymbol{0}$. The null and alternative hypotheses are

$$\mathbb{H}_0 : \mathbb{E}(\boldsymbol{z}_{bi} e_i) = \boldsymbol{0}$$
$$\mathbb{H}_1 : \mathbb{E}(\boldsymbol{z}_{bi} e_i) \neq \boldsymbol{0}.$$

That is, the null hypothesis is that the full set of moment conditions are valid, while the alternative hypothesis is that the instrument subset $\boldsymbol{z}_{bi}$ is correlated with e_i and thus an invalid instrument. Rejection of $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ is then interpreted as evidence that $\boldsymbol{z}_{bi}$ is misspecified as an instrument.

Based on the same reasoning as described in the previous section, to test $\mathbb{H}_0$ against $\mathbb{H}_1$ we consider a partitioned version of the regression (11.69)

$$e_i = \boldsymbol{z}_{ai}'\boldsymbol{\alpha}_a + \boldsymbol{z}_{bi}'\boldsymbol{\alpha}_b + \varepsilon_i$$

but now focus on the coefficient $\boldsymbol{\alpha}_b$. Given $\mathbb{E}(\boldsymbol{z}_{ai}e_i) = \mathbf{0}$, $\mathbb{H}_0$ is equivalent to $\boldsymbol{\alpha}_b = \mathbf{0}$. The equation is estimated by least-squares, replacing the unobseved e_i with the 2SLS residual $\widehat{e}_i$. The estimate of $\boldsymbol{\alpha}_b$ is

$$\widehat{\boldsymbol{\alpha}}_b = \left(\boldsymbol{Z}_b'\boldsymbol{M}_a\boldsymbol{Z}_b\right)^{-1}\boldsymbol{Z}_b'\boldsymbol{M}_a\widehat{\boldsymbol{e}}$$

where $\boldsymbol{M}_a = \boldsymbol{I}_n - \boldsymbol{Z}_a\left(\boldsymbol{Z}_a'\boldsymbol{Z}_a\right)^{-1}\boldsymbol{Z}_a'$. Newey (1985) showed that an optimal (asymptotically most powerful) test of $\mathbb{H}_0$ against $\mathbb{H}_1$ is to reject for large values of the score statistic

$$\begin{aligned} N &= \widehat{\boldsymbol{\alpha}}_b'\left(\widehat{\operatorname{var}(\widehat{\boldsymbol{\alpha}}_b)}\right)^{-}\widehat{\boldsymbol{\alpha}}_b \\ &= \frac{\widehat{\boldsymbol{e}}'\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{R} - \boldsymbol{R}'\widehat{\boldsymbol{X}}\left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widehat{\boldsymbol{e}}}{\widehat{\sigma}^2} \end{aligned}$$

where $\widehat{\boldsymbol{X}} = \boldsymbol{P}\boldsymbol{X}$, $\boldsymbol{P} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$, $\boldsymbol{R} = \boldsymbol{M}_a\boldsymbol{Z}_b$, and $\widehat{\sigma}^2 = \frac{1}{n}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$.

Independently from Newey (1985), Eichenbaum, Hansen, and Singleton (1988) proposed a test based on the difference of Sargan statistics. Letting S be the Sargan test statistic (11.70) based on the full instrument set and S_a be the Sargan test based on the instrument set $\boldsymbol{z}_{ai}$, the Sargan difference statistic is

$$C = S - S_a.$$

Specifically, let $\widetilde{\boldsymbol{\beta}}_{2\text{sls}}$ be the 2SLS estimator using the instruments $\boldsymbol{z}_{ai}$ only, set $\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}_{2\text{sls}}$, and set $\widetilde{\sigma}^2 = \frac{1}{n}\widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}}$. Then

$$S_a = \frac{\widetilde{\boldsymbol{e}}'\boldsymbol{Z}_a\left(\boldsymbol{Z}_a'\boldsymbol{Z}_a\right)^{-1}\boldsymbol{Z}_a'\widetilde{\boldsymbol{e}}}{\widetilde{\sigma}^2}.$$

An advantage of the C statistic is that it is quite simple to calculate from the standard regression output.

At this point it is useful to reflect on our stated requirement that $\ell_a > k$. Indeed, if $\ell_a < k$ then $\boldsymbol{z}_{ai}$ fails the order condition for identification and $\widetilde{\boldsymbol{\beta}}_{2\text{sls}}$ cannot be calculated. Thus $\ell_a \geq k$ is necessary to compute S_a and hence S. Furthermore, if $\ell_a = k$ then $\boldsymbol{z}_{ai}$ is just identified so while $\widetilde{\boldsymbol{\beta}}_{2\text{sls}}$ can be calculated, the statistic $S_a = 0$ so $C = S$. Thus when $\ell_a = k$ the subset test equals the full overidentification test so there is no gain from considering subset tests.

The C statistic S_a is asymptotically equivalent to replacing $\widetilde{\sigma}^2$ in S_a with $\widehat{\sigma}^2$, yielding the statistic

$$C^* = \frac{\widetilde{\boldsymbol{e}}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}}}{\widehat{\sigma}^2} - \frac{\widetilde{\boldsymbol{e}}'\boldsymbol{Z}_a\left(\boldsymbol{Z}_a'\boldsymbol{Z}_a\right)^{-1}\boldsymbol{Z}_a'\widetilde{\boldsymbol{e}}}{\widehat{\sigma}^2}.$$

It turns out that this is Newey's statistic N. These tests have chi-square asymptotic distributions.

Let c satisfy $\alpha = 1 - G_{\ell_b}(c)$.

Theorem 11.28.1 *Algebraically, $N = C^*$. Under Assumption 11.14.1 and $\mathbb{E}(e_i^2|\boldsymbol{z}_i) = \sigma^2$, as $n \to \infty$, $N \xrightarrow{d} \chi^2_{\ell_b}$ and $C \xrightarrow{d} \chi^2_{\ell_b}$. Thus the tests "Reject $\mathbb{H}_0$ if $N > c$" and "Reject $\mathbb{H}_0$ if $C > c$" are asymptotically equivalent and have asymptotic size α.*

Theorem 11.28.1 shows that N and C^* are identical, and are near equivalents to the convenient statistic C^*, and the appropriate asymptotic distribution is $\chi^2_{\ell_b}$. Computationally, the easiest method to implement a subset overidentification test is to estimate the model twice by 2SLS, first using the full instrument set $\boldsymbol{z}_i$ and the second using the partial instrument set $\boldsymbol{z}_{ai}$. Compute the Sargan statistics for both 2SLS regressions, and compute C as the difference in the Sargan statistics. In Stata, for example, this is simple to implement with a few lines of code.

We illustrate using the Card college proximity example. Our reported 2SLS estimates have $\ell - k = 1$ so there is no role for a subset overidentification test. (Recall, the number of overidentifying restrictions must exceed one.) To illustrate we consider adding extra instruments to the estimates in column 5 of Table 1.1 (the 2SLS estimates using *public*, *private*, *age*, and age^2 as instruments for *education*, *experience*, and $experience^2/100$). We add two instruments: the years of education of the *father* and the *mother* of the worker. These variables had been used in the earlier labor economics literature as instruments, but Card did not. (He used them as regression controls in some specifications.) The motivation for using parent's education as instruments is the hypothesis that parental education influences children's educational attainment, but does not directly influence their ability. The more modern labor economics literature has disputed this idea, arguing that children are educated in part at home, and thus parent's education has a direct impact on the skill attainment of children (and not just an indirect impact via educational attainment). The older view was that parent's education is a valid instrument, the modern view is that it is not valid. We can test this dispute using a overidentification subset test.

We do this by estimating the wage equation by 2SLS using *public*, *private*, *age*, age^2, *father*, and *mother*, as instruments for *education*, *experience*, and $experience^2/100$). We do not report the parameter estimates here, but observe that this model is overidentified with 3 overidentifying restrictions. We calculate the Sargan overidentification statistic. It is 7.9 with an asymptotic p-value (calculated using χ^2_3) of 0.048. This is a mild rejection of the null hypothesis of correct specification. As we argued in the previous section, this by itself is not reason to reject the model. Now we consider a subset overidentification test. We are interested in testing the validity of the two instruments *father* and *mother*, not the instruments *public*, *private*, *age*, age^2. To test the hypothesis that these two instruments are uncorrelated with the structural error, we compute the difference in Sargan statistic, $C = 7.9 - 0.5 = 7.4$, which has a p-value (calculated using χ^2_2) of 0.025. This is marginally statistically significant, meaning that there is evidence that *father* and *mother* are not valid instruments for the wage equation. Since the p-value is not smaller than 1%, it is not overwhelming evidence, but it still supports Card's decision to not use parental education as instruments for the wage equation.

We now prove the results in Theorem 11.28.1.

We first show that $N = C^*$. Define $\boldsymbol{P}_a = \boldsymbol{Z}_a \left(\boldsymbol{Z}_a'\boldsymbol{Z}_a\right)^{-1} \boldsymbol{Z}_a'$ and $\boldsymbol{P}_{\boldsymbol{R}} = \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{R}\right)^{-1}\boldsymbol{R}'$. Since $[\boldsymbol{Z}_a, \boldsymbol{R}]$ span $\boldsymbol{Z}$ we find $\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{R}} + \boldsymbol{P}_a$ and $\boldsymbol{P}_{\boldsymbol{R}}\boldsymbol{P}_a = \boldsymbol{0}$. It will be useful to note that

$$\boldsymbol{P}_{\boldsymbol{R}}\widehat{\boldsymbol{X}} = \boldsymbol{P}_{\boldsymbol{R}}\boldsymbol{P}\boldsymbol{X} = \boldsymbol{P}_{\boldsymbol{R}}\boldsymbol{X}$$
$$\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}} - \widehat{\boldsymbol{X}}'\boldsymbol{P}_{\boldsymbol{R}}\widehat{\boldsymbol{X}} = \boldsymbol{X}'\left(\boldsymbol{P} - \boldsymbol{P}_{\boldsymbol{R}}\right)\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{P}_a\boldsymbol{X}.$$

The fact that $\boldsymbol{X}'\boldsymbol{P}\widehat{\boldsymbol{e}} = \widehat{\boldsymbol{X}}'\widehat{\boldsymbol{e}} = \boldsymbol{0} =$ implies $\boldsymbol{X}'\boldsymbol{P}_{\boldsymbol{R}}\widehat{\boldsymbol{e}} = -\boldsymbol{X}'\boldsymbol{P}_a\widehat{\boldsymbol{e}}$. Finally, since $\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{e}}$,

$$\widetilde{\boldsymbol{e}} = \left(\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{P}_a\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_a\right)\widehat{\boldsymbol{e}}$$

so

$$\widetilde{\boldsymbol{e}}'\boldsymbol{P}_a\widetilde{\boldsymbol{e}} = \widehat{\boldsymbol{e}}'\left(\boldsymbol{P}_a - \boldsymbol{P}_a\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{P}_a\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_a\right)\widehat{\boldsymbol{e}}.$$

Applying the Woodbury matrix equality to the definition of N, and the above algebraic relationships,

$$\begin{aligned}
N &= \frac{\widehat{\boldsymbol{e}}'\boldsymbol{P}_{\boldsymbol{R}}\widehat{\boldsymbol{e}} + \widehat{\boldsymbol{e}}'\boldsymbol{P}_{\boldsymbol{R}}\widehat{\boldsymbol{X}}\left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}} - \widehat{\boldsymbol{X}}'\boldsymbol{P}_{\boldsymbol{R}}\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{P}_{\boldsymbol{R}}\widehat{\boldsymbol{e}}}{\widehat{\sigma}^2} \\
&= \frac{\widehat{\boldsymbol{e}}'\boldsymbol{P}\widehat{\boldsymbol{e}} - \widehat{\boldsymbol{e}}'\boldsymbol{P}_a\widehat{\boldsymbol{e}} + \widehat{\boldsymbol{e}}'\boldsymbol{P}_a\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{P}_a\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}_a\widehat{\boldsymbol{e}}}{\widehat{\sigma}^2} \\
&= \frac{\widehat{\boldsymbol{e}}'\boldsymbol{P}\widehat{\boldsymbol{e}} - \widetilde{\boldsymbol{e}}'\boldsymbol{P}_a\widetilde{\boldsymbol{e}}}{\widehat{\sigma}^2} \\
&= C^*
\end{aligned}$$

as claimed.

We next establish the asymptotic distribution. Since $\boldsymbol{Z}_a$ is a subset of $\boldsymbol{Z}$, $\boldsymbol{P}\boldsymbol{M}_a = \boldsymbol{M}_a\boldsymbol{P}$, thus $\boldsymbol{P}\boldsymbol{R} = \boldsymbol{R}$ and $\boldsymbol{R}'\boldsymbol{X} = \boldsymbol{R}'\widehat{\boldsymbol{X}}$. Consequently

$$\begin{aligned}\frac{1}{\sqrt{n}}\boldsymbol{R}'\widehat{\boldsymbol{e}} &= \frac{1}{\sqrt{n}}\boldsymbol{R}'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) \\ &= \frac{1}{\sqrt{n}}\boldsymbol{R}'\left(\boldsymbol{I}_n - \boldsymbol{X}\left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\right)\boldsymbol{e} \\ &= \frac{1}{\sqrt{n}}\boldsymbol{R}'\left(\boldsymbol{I}_n - \widehat{\boldsymbol{X}}\left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\right)\boldsymbol{e}. \\ &\xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_2\right)\end{aligned}$$

where

$$\boldsymbol{V}_2 = \operatorname*{plim}_{n\to\infty}\left(\frac{1}{n}\boldsymbol{R}'\boldsymbol{R} - \frac{1}{n}\boldsymbol{R}'\widehat{\boldsymbol{X}}\left(\frac{1}{n}\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\frac{1}{n}\widehat{\boldsymbol{X}}'\boldsymbol{R}\right).$$

It follows that $N = C^* \xrightarrow{d} \chi^2_{\ell_b}$ as claimed. Since $C = C^* + o_p(1)$ it has the same limiting distribution.

11.29 Local Average Treatment Effects

In a pair of influential papers, Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996) proposed an new interpretation of the instrumental variables estimator using the potential outcomes model introduced in Section 2.29.

We will restrict attention to the case that the endogenous regressor x and excluded instrument z are binary variables. We write the model as a pair of potential outcome functions. The dependent variable y is a function of the regressor and an unobservable vector $\boldsymbol{u}$

$$y = h\left(x, \boldsymbol{u}\right)$$

and the endogenous regressor x is a function of the instrument z and $\boldsymbol{u}$

$$x = g\left(z, \boldsymbol{u}\right).$$

By specifying $\boldsymbol{u}$ as a vector there is no loss of generality in letting both equations depend on $\boldsymbol{u}$.

In this framework, the outcomes are determined by the random vector $\boldsymbol{u}$ and the exogenous instrument z. This determines x, which determines y. To put this in the context of the college proximity example, the variable $\boldsymbol{u}$ is everything specific about an individual. Given college proximity z, the person decides to attend college or not. The person's wage is determined by the individual attributes $\boldsymbol{u}$ as well as college attendence x, but is not directly affected by college proximity z.

We can omit the random variable $\boldsymbol{u}$ from the notation as follows. An individual i has a realization $\boldsymbol{u}_i$. We then set $y_i(x) = h\left(x, \boldsymbol{u}_i\right)$ and $x_i(z) = g\left(z, \boldsymbol{u}_i\right)$. Also, given a realization z_i the observables are $x_i = x_i(z_i)$ and $y_i = y_i(x_i)$.

In this model the causal effect of college is for individual i is

$$C_i = y_i(1) - y_i(0).$$

As discussed in Section 2.29, in general this is individual-specific.

We would like to learn about the distribution of the causal effects, or at least features of the distribution. A common feature of interest is the average treatment effect (ATE)

$$ATE = \mathbb{E}\left(C_i\right) = \mathbb{E}\left(y_i(1) - y_i(0)\right).$$

This, however, it typically not feasible to estimate allowing for endogenous x without strong assumptions (such as that the causal effect C_i is constant across individuals). The treatment effect literature has explored what features of the distribution of C_i can be estimated.

One particular feature of interest, and emphasized by Imbens and Angrist (1994), is known as the local average treatment effect (LATE), and is roughly the average effect upon those effected by the instrumental variable. To understand LATE, it is helpful to consider the college proximity example using the potential outcomes framework. In this framework, each person is fully characterized by their individual unobservable $\boldsymbol{u}_i$. Given $\boldsymbol{u}_i$, their decision to attend college is a function of the proximity indicator z_i. For some students, proximity has no effect on their decision. For other students, it has an effect in the specific sense that given $z_i = 1$ they choose to attend college while if $z_i = 0$ they choose to not attend. We can summarize the possibilites with the following chart, which is based on labels developed by Angrist, Imbens and Rubin (1996).

	$x(0) = 0$	$x(0) = 1$
$x(1) = 0$	Never Takers	Deniers
$x(1) = 1$	Compliers	Always Takers

The columns indicate the college attendence decision given $z = 0$. The rows indicate the college attendence decision given $z = 1$. The four entries are labels given four types of individuals based on these decisions. The upper-left entry are the individuals who do not attend college regardless of z. They are called "Never Takers". The lower-right entry are the individuals who conversely attend college regardless of z. They are called "Always Takers". The bottom left are the individuals who only attend college if they live close to one. They are called "Compliers". The upper right entry is a bit of a challenge. These are individuals who attend college only if they do not live close to one. They are called "Deniers". Imbens and Angrist discovered that to identify the parameters of interest we need to assume that there are no Deniers, or equivalently that $x(1) \geq x(0)$, which they label as a "monotonicity" condition – that increasing the instrument cannot decrease x for any individual.

We can distinguish the types in the table by the relative values of $x(1) - x(0)$. For Never-Takers and Always-Takers, $x(1) - x(0) = 0$, while for Deniers, $x(1) - x(0) = 1$

We are interested in the causal effect $C_i = h(1, \boldsymbol{u}) - h(0, \boldsymbol{u})$ of college attendence on wages. Consider the average causal effect among the different types. Among Never-Takers and Always-Takers, $x(1) = x(0)$ so

$$\mathbb{E}\left(y_i(1) - y_i(0) | x(1) = x(0)\right)$$

Suppose we try and estimate its average value, conditional for each the three types of individuals: Never-Takers, Always-Takers, and Compliers. It would impossible for the Never-Takers and Always-Takers. For the former, none attend college so it would be impossible to ascertain the effect of college attendence, and similarly for the latter since they all attend college. Thus the only group for which we can estimate the average causal effect are the Compliers. This is

$$\text{LATE} = \mathbb{E}\left(y_i(1) - y_i(0) | x_i(1) > x_i(0)\right).$$

Imbens and Angrist called this the **local average treatment effect (LATE)** as it is the average treatment effect for the sub-population whose endogenous regressor is affected by changes in the instrumental variable.

Interestingly, we show below that

$$\text{LATE} = \frac{\mathbb{E}\left(y_i \mid z_i = 1\right) - \mathbb{E}\left(y_i \mid z_i = 0\right)}{\mathbb{E}\left(x_i \mid z_i = 1\right) - \mathbb{E}\left(x_i \mid z_i = 0\right)}. \tag{11.71}$$

That is, LATE equals the Wald expression (11.32) for the slope coefficient in the IV regression model. This means that the standard IV estimator is an estimator of LATE. Thus when treatment effects are potentially heterogeneous, we can interpret IV as an estimator of LATE. The equality (11.71) occurs under the following conditions.

Assumption 11.29.1 *$\boldsymbol{u}_i$ and z_i are independent; and* $\Pr(x_i(1) - x_i(0) < 0) = 0$.

One interesting feature about LATE is that its value can depend on the instrument z_i and the distribution of causal effects C_i in the population. To make this concrete, suppose that instead of the Card proximity instrument, we consider an instrument based on the financial cost of local college attendence. It is reasonable to expect that while the set of students affected by these two instruments are similar, the two sets of students will not be the same. That is, some students may be responsive to proximity but not finances, and conversely. If the causal effect C_i has a different average in these two groups of students, then LATE will be different when calculated with these two instruments. Thus LATE can vary by the choice of instrument.

How can that be? How can a well-defined parameter depend on the choice of instrument? Doesn't this contradict the basic IV regression model? The answer is that the basic IV regression model is more restrictive – it specifies that the causal effect β is common across all individuals. Thus its value is the same regardless of the choice of specific instrument (so long as it satisfies the instrumental variables assumptions). In contrast, the potential outcomes framework is more general, allowing for the causal effect to vary across individuals. What this analysis shows us is that in this context is quite possible for the LATE coefficient to vary by instrument. This occurs when causal effects are heterogeneous.

One implication of the LATE framework is that IV estimates should be interpreted as causal effects only for the population of compliers. Interpretation should focus on the population of potential compliers and extension to other populations should be done with caution. For example, in the Card proximity model, the IV estimates of the causal return to schooling presented in Table 11.1 should be interpreted as applying to the population of students who are incentivized to attend college by the presence of a college within their home county. The estimates should not be applied to other students.

Formally, the analysis of this section examined the case of a binary instrument and endogenous regressor. How does this generalize? Suppose that the regressor x is discrete, taking $J+1$ discrete values. We can then rewrite the model as one with J binary endogenous regressors. If we then have J binary instruments, we are back in the Imbens-Angrist framework (assuming the instruments have a monotonic impact on the endogenous regressors). A benefit is that with a larger set of instruments it is plausible that the set of compliers in the population is expanded.

We close this section by showing (11.71) under Assumption 11.29.1. The realized value of x_i can be written as

$$x_i = (1 - z_i)\, x_i(0) + z_i x_i(1) = x_i(0) + z_i \left(x_i(1) - x_i(0)\right).$$

Similarly

$$y_i = y_i(0) + x_i \left(y_i(1) - y_i(0)\right) = y_i(0) + x_i C_i.$$

Combining,

$$y_i = y_i(0) + x_i(0) C_i + z_i \left(x_i(1) - x_i(0)\right) C_i.$$

The independence of $\boldsymbol{u}_i$ and z_i implies independence of $(y_i(0), y_i(1), x_i(0), x_i(1), C_i)$ and z_i. Thus

$$\mathbb{E}\left(y_i | z_i = 1\right) = \mathbb{E}\left(y_i(0)\right) + \mathbb{E}\left(x_i(0) C_i\right) + \mathbb{E}\left(\left(x_i(1) - x_i(0)\right) C_i\right)$$

and

$$\mathbb{E}\left(y_i | z_i = 0\right) = \mathbb{E}\left(y_i(0)\right) + \mathbb{E}\left(x_i(0) C_i\right).$$

Subtracting we obtain

$$\begin{aligned}
\mathbb{E}\left(y_i | z_i = 1\right) - \mathbb{E}\left(y_i | z_i = 0\right) &= \mathbb{E}\left(\left(x_i(1) - x_i(0)\right) C_i\right) \\
&= 1 \cdot \mathbb{E}\left(C_i | x_i(1) - x_i(0) = 1\right) \Pr\left(x_i(1) - x_i(0) = 1\right) \\
&+ 0 \cdot \mathbb{E}\left(C_i | x_i(1) - x_i(0) = 0\right) \Pr\left(x_i(1) - x_i(0) = 0\right) \\
&+ (-1) \cdot \mathbb{E}\left(C_i | x_i(1) - x_i(0) = -1\right) \Pr\left(x_i(1) - x_i(0) = -1\right) \\
&= \mathbb{E}\left(C_i | x_i(1) - x_i(0) = 1\right) \left(\mathbb{E}\left(x_i \mid z_i = 1\right) - \mathbb{E}\left(x_i \mid z_i = 0\right)\right)
\end{aligned}$$

where the final equality uses $\Pr(x_i(1) - x_i(0) < 0) = 0$ and

$$\Pr(x_i(1) - x_i(0) = 1) = \mathbb{E}(x_i(1) - x_i(0)) = \mathbb{E}(x_i \mid z_i = 1) - \mathbb{E}(x_i \mid z_i = 0).$$

Rearranging

$$\text{LATE} = \mathbb{E}(C_i | x_i(1) - x_i(0) = 1) = \frac{\mathbb{E}(y_i | z_i = 1) - \mathbb{E}(y_i | z_i = 0)}{\mathbb{E}(x_i \mid z_i = 1) - \mathbb{E}(x_i \mid z_i = 0)}$$

as claimed.

11.30 Identification Failure

Recall the reduced form equation

$$\boldsymbol{x}_{2i} = \boldsymbol{\Gamma}_{12}'\boldsymbol{z}_{1i} + \boldsymbol{\Gamma}_{22}'\boldsymbol{z}_{2i} + \boldsymbol{u}_{2i}.$$

The parameter $\boldsymbol{\beta}$ fails to be identified if $\boldsymbol{\Gamma}_{22}$ has deficient rank. The consequences of identification failure for inference are quite severe.

Take the simplest case where $k_1 = 0$ and $k_2 = \ell_2 = 1$. Then the model may be written as

$$\begin{aligned} y_i &= x_i\beta + e_i \\ x_i &= z_i\gamma + u_i \end{aligned} \tag{11.72}$$

and $\Gamma_{22} = \gamma = \mathbb{E}(z_i x_i) / \mathbb{E}\left(z_i^2\right)$. We see that β is identified if and only if $\gamma \neq 0$, which occurs when $\mathbb{E}(x_i z_i) \neq 0$. Thus identification hinges on the existence of correlation between the excluded exogenous variable and the included endogenous variable.

Suppose this condition fails. In this case $\gamma = 0$ and $\mathbb{E}(x_i z_i) = 0$. We now analyze the distribution of the least-squares and IV estimators of β. For simplicity we assume conditional homoskedasticity and normalize the variances to unity. Thus

$$\begin{aligned} \operatorname{var}\left(\begin{pmatrix} e_i \\ u_i \end{pmatrix} \mid z_i\right) &= \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \\ \mathbb{E}\left(z_i^2\right) &= 1. \end{aligned} \tag{11.73}$$

The errors have non-zero correlation $\rho \neq 0$ which occurs when the variables are endogenous.

By the CLT we have the joint convergence

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} z_i e_i \\ z_i u_i \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim \mathrm{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \tag{11.74}$$

It is convenient to define $\xi_0 = \xi_1 - \rho\xi_2$ which is normal and independent of ξ_2.

As a benchmark, it is useful to observe that the least-squares estimator of β satisfies

$$\widehat{\beta}_{\text{ols}} - \beta = \frac{n^{-1}\sum_{i=1}^{n} u_i e_i}{n^{-1}\sum_{i=1}^{n} u_i^2} \xrightarrow{p} \rho \neq 0 \tag{11.75}$$

so endogeneity causes $\widehat{\beta}_{\text{ols}}$ to be inconsistent for β.

Under identification failure $\gamma = 0$ the asymptotic distribution of the IV estimator is

$$\widehat{\beta}_{\text{iv}} - \beta = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i e_i}{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i x_i} \xrightarrow{d} \frac{\xi_1}{\xi_2} = \rho + \frac{\xi_0}{\xi_2}.$$

This asymptotic convergence result uses the continuous mapping theorem, which applies since the function ξ_1/ξ_2 is continuous everywhere except at $\xi_2 = 0$, which occurs with probability equal to zero.

This limiting distribution has several notable features.

First, $\widehat{\beta}_{\text{iv}}$ does not converge in probability to a limit, rather it converges in distribution to a random variable. Thus the IV estimator is inconsistent. Indeed, it is not possible to consistently estimate an unidentified parameter and β is not identified when $\gamma = 0$.

Second, the ratio ξ_0/ξ_2 is symmetrically distributed about zero, so the median of the limiting distribution of $\widehat{\beta}_{\text{iv}}$ is $\beta + \rho$. This means that the IV estimator is median biased under endogeneity. Thus under identification failure the IV estimator does not correct the centering (median bias) of least-squares.

Third, the ratio ξ_0/ξ_2 of two independent normal random variables is Cauchy distributed. This is particularly nasty, as the Cauchy distribution does not have a finite mean. The distribution has thick tails meaning that extreme values occur with higher frequency than the normal, and inferences based on the normal distribution can be quite incorrect.

Together, these results show that $\gamma = 0$ renders the IV estimator particularly poorly behaved – it is inconsistent, median biased, and non-normally distributed.

We can also examine the behavior of the t-statistic. For simplicity consider the classical (homoskedastic) t-statistic. The error variance estimate has the asymptotic distribution

$$\begin{aligned}
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\widehat{\beta}_{\text{iv}}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} e_i^2 - \frac{2}{n}\sum_{i=1}^{n} e_i x_i\left(\widehat{\beta}_{\text{iv}} - \beta\right) + \frac{1}{n}\sum_{i=1}^{n} x_i^2\left(\widehat{\beta}_{\text{iv}} - \beta\right)^2 \\
&\xrightarrow{d} 1 - 2\rho\frac{\xi_1}{\xi_2} + \left(\frac{\xi_1}{\xi_2}\right)^2.
\end{aligned}$$

Thus the t-statistic has the asymptotic distribution

$$T = \frac{\widehat{\beta}_{\text{iv}} - \beta}{\sqrt{\widehat{\sigma}^2 \sum_{i=1}^{n} z_i^2 / \left|\sum_{i=1}^{n} z_i x_i\right|}} \xrightarrow{d} \frac{\xi_1/\xi_2}{\sqrt{1 - 2\rho\frac{\xi_1}{\xi_2} + \left(\frac{\xi_1}{\xi_2}\right)^2}}.$$

The limiting distribution is non-normal, meaning that inference using the normal distribution will be (considerably) incorrect. This distribution depends on the correlation ρ. The distortion from the normal is increasing in ρ. Indeed as $\rho \to 1$ we have $\xi_1/\xi_2 \to_p 1$ and the unexpected finding $\widehat{\sigma}^2 \to_p 0$. The latter means that the conventional standard error $s(\widehat{\beta}_{\text{iv}})$ for $\widehat{\beta}_{\text{iv}}$ also converges in probability to zero. This implies that the t-statistic diverges in the sense $|T| \to_p \infty$. In this situations users may incorrectly interpret estimates as precise, despite the fact that they are useless.

11.31 Weak Instruments

In the previous section we examined the extreme consequences of full identification failure. Unfortunately many of the same problems extend to the context where identification is weak in the sense that the reduced form coefficient matrix $\mathbf{\Gamma}_{22}$ is full rank but small.

A rich asymptotic distribution theory has been developed to understand this setting by modeling $\mathbf{\Gamma}_{22}$ as "local-to-zero". The seminal contributions are Staiger and Stock (1997) and Stock and Yogo (2005). The theory was extended to nonlinear GMM estimation by Stock and Wright (2000).

In this section we focus exclusively on the case of one right-hand-side endogenous variable ($k_2 = 1$). We consider the case of multiple endogenous variables in the next section. Our general theory will allow for any arbitrary number of instruments and regressors, but for the sake of clear

exposition we will focus on the very simple case of no included exogenous variables ($k_1 = 0$) and just one exogenous instrument ($\ell_2 = 1$), which is model (11.72) from the previous section

$$\begin{aligned} y_i &= x_i\beta + e_i \\ x_i &= z_i\gamma + u_i. \end{aligned}$$

Furthermore, as in Section 11.30 we assume conditional homoskedasticity and normalize the variances as in (11.73).

The question of primary interest is to determine conditions on the reduced form under which the IV estimator of the structural equation is well behaved, and secondly, what statistical tests can be used to learn if these conditions are satisfied.

In Section 11.30 we assumed complete identification failure in the sense that $\gamma = 0$. We now want to assume that identification does not completely fail, but is weak in the sense that γ is small. The technical device which yields a useful distributional theory is to assume that the reduced form parameter is **local-to-zero**, specifically

$$\gamma = n^{-1/2}\mu \tag{11.76}$$

where μ is a free parameter. The $n^{-1/2}$ scaling is picked because it provides just the right balance to allow a useful distribution theory. The local-to-zero assumption (11.76) is not meant to be taken literally but rather is meant to be a useful distributional approximation. The parameter μ indexes the degree of identification. Larger $|\mu|$ implies stronger identification; smaller $|\mu|$ implies weaker identification.

We now derive the asymptotic distribution of the least-squares and IV estimators under the local-to-unity assumption (11.76).

First, the least-squares estimator satisfies

$$\widehat{\beta}_{\text{ols}} - \beta = \frac{n^{-1}\sum_{i=1}^n x_i e_i}{n^{-1}\sum_{i=1}^n x_i^2} = \frac{n^{-1}\sum_{i=1}^n u_i e_i}{n^{-1}\sum_{i=1}^n u_i^2} + o_p(1) \xrightarrow{p} \rho \neq 0$$

which is the same as in (11.75). Thus the least-squares estimator is inconsistent for β under endogeneity.

Second, we derive the distribution of the IV estimator. The joint convergence (11.74) holds, and the local-to-zero assumption implies

$$\begin{aligned} \frac{1}{\sqrt{n}}\sum_{i=1}^n z_i x_i &= \frac{1}{\sqrt{n}}\sum_{i=1}^n z_i^2\gamma + \frac{1}{\sqrt{n}}\sum_{i=1}^n z_i u_i \\ &= \frac{1}{n}\sum_{i=1}^n z_i^2\mu + \frac{1}{\sqrt{n}}\sum_{i=1}^n z_i u_i \\ &\xrightarrow{d} \mu + \xi_2. \end{aligned}$$

This allows us to calculate the asymptotic distribution of the IV estimator.

$$\widehat{\beta}_{\text{ols}} - \beta = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i e_i}{\frac{1}{\sqrt{n}}\sum_{i=1}^n z_i x_i} \xrightarrow{d} \frac{\xi_1}{\mu + \xi_2}.$$

This asymptotic convergence result uses the continuous mapping theorem, which applies since the function $\xi_1/(\mu + \xi_2)$ is a continuous function everywhere except at $\xi_2 = -\mu$, which occurs with probability equal to zero.

As in the case of complete identification failure, we find that $\widehat{\beta}_{\text{iv}}$ is inconsistent for β and its asymptotic distribution is non-normal. The distortion is affected by the coefficient μ. As $\mu \to \infty$

the distribution converges in probability to zero, meaning that $\widehat{\beta}_{\text{iv}}$ is consistent for β. This is the classic "strong identification" context.

We also examine the behavior of the classical (homoskedastic) t-statistic for the IV estimator. Note

$$\begin{aligned}\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\widehat{\beta}_{\text{iv}}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 - \frac{2}{n}\sum_{i=1}^{n} e_i x_i\left(\widehat{\beta}_{\text{iv}} - \beta\right) + \frac{1}{n}\sum_{i=1}^{n} x_i^2\left(\widehat{\beta}_{\text{iv}} - \beta\right)^2 \\ &\xrightarrow{d} 1 - 2\rho\frac{\xi_1}{\mu+\xi_2} + \left(\frac{\xi_1}{\mu+\xi_2}\right)^2.\end{aligned}$$

Thus

$$T = \frac{\widehat{\beta}_{\text{iv}} - \beta}{\sqrt{\widehat{\sigma}^2\sum_{i=1}^{n} z_i^2}/\left|\sum_{i=1}^{n} z_i x_i\right|} \xrightarrow{d} \frac{\xi_1}{\sqrt{1 - 2\rho\frac{\xi_1}{\mu+\xi_2} + \left(\frac{\xi_1}{\mu+\xi_2}\right)^2}} \stackrel{def}{=} S. \tag{11.77}$$

In general, S is non-normal, and its distribution depends on the parameters ρ and μ.

Can we use the distribution S for inference on β? The distribution depends on two unknown parameters, and neither is consistently estimable. (Thus we cannot simply use the distribution in (11.77) with ρ and μ replaced with estimates.) To eliminate the dependence on ρ one possibility is to use the "worst case" value, which turns out to be $\rho = 1$. By worst-case we mean that value which causes the greatest distortion away from normal critical values. Setting $\rho = 1$ we have the considerable simplification

$$S = S_1 = \xi\left|1 + \frac{\xi}{\mu}\right| \tag{11.78}$$

where $\xi \sim \mathrm{N}(0,1)$. When the model is strongly identified (so $|\mu|$ is very large) then $S_1 \approx \xi$ is standard normal, consistent with classical theory. However when $|\mu|$ is very small (but non-zero) $|S_1| \approx \xi^2/\mu$ (in the sense that this term dominates), which is a scaled χ_1^2 and quite far from normal. As $|\mu| \to 0$ we find the extreme case $|S_1| \to_p \infty$.

While (11.78) is a convenient simplification it does not yield a useful approximation for inference since the distribution in (11.78) is highly dependent on the unknown μ. If we try to take the worst-case value of μ, which is $\mu = 0$, we find that $|S_1|$ diverges and all distributional approximations fail.

To break this impasse, Stock and Yogo (2005) recommended a constructive alternative. Rather than using the worst-case μ, they suggested finding a threshold such that if μ exceeds this threshold then the distribution (11.78) is not "too badly" distorted from the normal distribuiton.

Specifically, the Stock-Yogo recommendation can be summarized by two steps. First, the distribution result (11.78) can be used to find a threshold value τ^2 such that if $\mu^2 \geq \tau^2$ then the size of the nominal[1] 5% test "Reject if $|T| \geq 1.96$" has asymptotic size $\Pr\left(|S_1| \geq 1.96\right) \leq 0.15$. This means that while the goal is to obtain a test with size 5%, we recognize that there may be size distortion due to weak instruments and are willing to tolerate a specific size distortion, for example 10% distortion (allow for actual size up to 15%, or more generally r). Second, they use the asymptotic distribution of the reduced-form (first stage) F statistic to test if the actual unknown value of μ^2 exceeds the threshold τ^2. These two steps together give rise to the rule-of-thumb that the first-stage F statistic should exceed 10 in order to achieve reliable IV inference. (This is for the case of one instrumental variable. If there is more than one instrument then the rule-of-thumb changes.) We now describe the steps behind this reasoning in more detail.

[1] The term "nominal size" of a test is the official intended size – the size which would obtain under ideal circumstances. In this context the test "Reject if $|T| \geq 1.96$" has nominal size 0.05 as this would be the asymptotic rejection probability in the ideal context of strong instruments.

The first step is to use the distribution (11.77) to determine the threshold τ^2. Formally, the goal is to find the value of $\tau^2 = \mu^2$ at which the asymptotic size of a nominal 5% test is actually r (e.g. $r = 0.15$)

$$\Pr\left(|S_1| \geq 1.96\right) \leq r.$$

By some algebra and using the quadratic formula the event $|\xi\left(1 + \xi/\mu\right)| < x$ is the same as

$$\frac{\mu^2}{4} - x\mu < \left(\xi + \frac{\mu}{2}\right)^2 < \frac{\mu^2}{4} + x\mu.$$

The random variable between the inequalities is distributed $\chi^2_1(\mu^2/4)$, a noncentral chi-square with one degree of freedom and noncentrality parameter $\mu^2/4$. Thus

$$\begin{aligned}\Pr\left(|S_1| \geq x\right) &= \Pr\left(\chi^2_1\left(\frac{\mu^2}{4}\right) \geq \frac{\mu^2}{4} + x\mu\right) + \Pr\left(\chi^2_1\left(\frac{\mu^2}{4}\right) \leq \frac{\mu^2}{4} - x\mu\right) \\ &= 1 - G\left(\frac{\mu^2}{4} + x\mu, \frac{\mu^2}{4}\right) + G\left(\frac{\mu^2}{4} - x\mu, \frac{\mu^2}{4}\right) \end{aligned} \tag{11.79}$$

where $G\left(u, \lambda\right)$ is the distribution function of $\chi^2_1(\lambda)$. Hence the desired threshold τ^2 solves

$$1 - G\left(\frac{\tau^2}{4} + 1.96\tau, \frac{\tau^2}{4}\right) + G\left(\frac{\tau^2}{4} - 1.96\tau, \frac{\tau^2}{4}\right) = r$$

or effectively

$$G\left(\frac{\tau^2}{4} + 1.96\tau, \frac{\tau^2}{4}\right) = 1 - r$$

since $\tau^2/4 - 1.96\tau < 0$ for relevant values of τ. The numerical solution (computed with the non-central chi-square distribution function, e.g. `ncx2cdf` in MATLAB) is $\tau^2 = 1.70$ when $r = 0.15$. (That is, the command `ncx2cdf(1.7/4+1.96*sqrt(1.7),1,1.7/4)` yields the answer `0.8500`. Stock and Yogo (2005) approximate the same calculation using simulation methods and report $\tau^2 = 1.82$.)

This calculation means that if the true reduced form coefficient satisfies $\mu^2 \geq 1.7$, or equivalently if $\gamma^2 \geq 1.7/n$, then the (asymptotic) size of a nominal 5% test on the structural parameter is no larger than 15%.

To summarize the Stock-Yogo first step, we calculate the minimum value τ^2 for μ^2 sufficient to ensure that the asymptotic size of a nominal 5% t-test does not exceed r, and find that $\tau^2 = 1.70$ for $r = 0.15$.

The Stock-Yogo second step is to find a critical value for the first-stage F statistic sufficient to reject the hypothesis that $\mathbb{H}_0 : \mu^2 = \tau^2$ against $\mathbb{H}_1 : \mu^2 > \tau^2$. We now describe this procedure.

They suggest testing $\mathbb{H}_0 : \mu^2 = \tau^2$ at the 5% size using the first stage F statistic. If the F statistic is small so that the test does not reject then we should be worried that the true value of μ^2 is small and there is a weak instrument problem. On the other hand if the F statistic is large so that the test rejects then we can have some confidence that the true value of μ^2 is sufficiently large that the weak instrument problem is not too severe.

To implement the test we need to calculate an appropriate critical value. It should be calculated under the null hypothesis $\mathbb{H}_0 : \mu^2 = \tau^2$. This is different from a conventional F test (which has the null hypothesis $\mathbb{H}_0 : \mu^2 = 0$).

We start by calculating the asymptotic distribution of F. Since there is just one regressor and one instrument in our simplified setting, the first-stage F statistic is the squared t-statistic from the reduced form, and given our previous calculations has the asymptotic distribution

$$F = \frac{\widehat{\gamma}^2}{s\left(\widehat{\gamma}\right)^2} = \frac{\left(\sum_{i=1}^n z_i x_i\right)^2}{\left(\sum_{i=1}^n x_i^2\right)\widehat{\sigma}_u^2} \xrightarrow{d} \left(\mu + \xi_2\right)^2 \sim \chi^2_1\left(\mu^2\right).$$

This is a non-central chi-square distribution with one degree of freedom and non-centrality parameter μ^2. The distribution function of the latter is $G(u, \mu^2)$.

To test $\mathbb{H}_0 : \mu^2 = \tau^2$ against $\mathbb{H}_1 : \mu^2 > \tau^2$ we reject for $F \geq c$ where c is selected so that the asymptotic rejection probability

$$\Pr(F \geq c) \to \Pr\left(\chi_1^2\left(\mu^2\right) \geq c\right) = 1 - G\left(c, \mu^2\right)$$

equals 0.05 under $\mathbb{H}_0 : \mu^2 = \tau^2$, or equivalently

$$G\left(c, \tau^2\right) = G(c, 1.7) = 0.95.$$

This can be found using the non-central chi-square quantile function, e.g. the function $Q(p, d)$ which solves $G(Q(p, d), d) = p$. We find that

$$c = Q(0.95, 1.7) = 8.7.$$

In MATLAB, this can be computed by `ncx2inv(.95,1.7)`. (Stock and Yogo (2005) report $c = 9.0$ since they used $\tau^2 = 1.82$.)

This means that if $F > 8.7$ we can reject $\mathbb{H}_0 : \mu^2 = 1.7$ against $\mathbb{H}_1 : \mu^2 > 1.7$ with an asymptotic 5% test. In this context we should expect the IV estimate and tests to be reasonably well behaved. However, if $F < 8.7$ then we should be cautious about the IV estimator, confidence intervals, and tests. This finding led Staiger and Stock (1997) to propose the informal "rule of thumb" that the first stage F statistic should exceed 10. Notice that F exceeding 8.7 (or 10) is equivalent to the reduced form t-statistic exceeding 2.94 (or 3.16), which is considerably larger than a conventional check if the t-statistic is "significant". Equivalently, the recommended rule-of-thumb for the case of a single instrument is to estimate the reduced form and verify that the t-statistic for exclusion of the instrumental variable exceeds 3 in absolute value.

Does the proposed procedure control the asymptotic size of a 2SLS test? The first step has asymptotic size bounded below r (e.g. 15%). The second step has asymptotic size 5%. By the Bonferroni bound (see Section 9.20) the two steps together have asymptotic size bounded below $r + 0.05$ (e.g. 20%). We can thus call the Stock-Yogo procedure a rigorous test with asymptotic size $r + 0.05$ (or 20%).

Our analysis has been confined to the case $k_2 = \ell_2 = 1$. Stock and Yogo (2005) also examine the case of $\ell_2 > 1$ (which requires numerical simulation to solve), and both the 2SLS and LIML estimators. They show that the F statistic critical values depend on the number of instruments ℓ_2 as well as the estimator. We report their calculations here.

F Statistic 5% Critical Value for Weak Instruments, $k_2 = 1$

	Maximal Size r							
	2SLS				LIML			
ℓ_2	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
1	16.4	9.0	6.7	5.5	16.4	9.0	6.7	5.5
2	19.9	11.6	8.7	7.2	8.7	5.3	4.4	3.9
3	22.3	12.8	9.5	7.8	6.5	4.4	3.7	3.3
4	24.6	14.0	10.3	8.3	5.4	3.9	3.3	3.0
5	26.9	15.1	11.0	8.8	4.8	3.6	3.0	2.8
6	29.2	16.2	11.7	9.4	4.4	3.3	2.9	2.6
7	31.5	17.4	12.5	9.9	4.2	3.2	2.7	2.5
8	33.8	18.5	13.2	10.5	4.0	3.0	2.6	2.4
9	36.2	19.7	14.0	11.1	3.8	2.9	2.5	2.3
10	38.5	20.9	14.8	11.6	3.7	2.8	2.5	2.2
15	50.4	26.8	18.7	12.2	3.3	2.5	2.2	2.0
20	62.3	32.8	22.7	17.6	3.2	2.3	2.1	1.9
25	74.2	38.8	26.7	20.6	3.8	2.2	2.0	1.8
30	86.2	44.8	30.7	23.6	3.9	2.2	1.9	1.7

One striking feature about these critical values is that those for the 2SLS estimator are strongly increasing in ℓ_2 while those for the LIML estimator are decreasing in ℓ_2. This means that when the number of instruments ℓ_2 is large, 2SLS requires a much stronger reduced form (larger μ^2) in order for inference to be reliable, but this is not the case for LIML. This is direct evidence that inference is less sensitive to weak instruments when estimation is by LIML rather than 2SLS. This makes a strong case for using LIML rather than 2SLS, especially when ℓ_2 is large or the instruments are potentially weak.

We now summarize the recommended Staiger-Stock/Stock-Yogo procedure for $k_1 \geq 1$, $k_2 = 1$, and $\ell_2 \geq 1$. The structural equation and reduced form equations are

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + x_{2i}\beta_2 + e_i$$
$$x_{2i} = \boldsymbol{x}_{1i}'\boldsymbol{\gamma}_1 + \boldsymbol{z}_{2i}'\boldsymbol{\gamma}_2 + u_i.$$

The reduced form is estimated by least-squares

$$x_{2i} = \boldsymbol{x}_{1i}'\widehat{\boldsymbol{\gamma}}_1 + \boldsymbol{z}_{2i}'\widehat{\boldsymbol{\gamma}}_2 + \widehat{u}_i$$

and the structural equation by either 2SLS or LIML:

$$y_i = \boldsymbol{x}_{1i}'\widehat{\boldsymbol{\beta}}_1 + x_{2i}\widehat{\beta}_2 + \widehat{e}_i.$$

Let F be the F statistic for $\mathbb{H}_0 : \boldsymbol{\gamma}_2 = 0$ in the reduced form equation. Let $s(\widehat{\beta}_2)$ be a standard error for β_2 in the structural equation. The procedure is:

1. Compare F with the critical values c in the above table, with the row selected to match the number of excluded instruments ℓ_2, and the columns to match the estimation method (2SLS or LIML) and the desired size r.

2. If $F > c$ then report the 2SLS or LIML estimates with conventional inference.

The Stock-Yogo test can be implemented in Stata using the command `estat firststage` after `ivregress 2sls` or `ivregres liml` if a standard (non-robust) covariance matrix has been specified (that is, without the '`,r`' option).

There are possible extensions to the Stock-Yogo procedure.

One modest extension is to use the information to convey the degree of confidence in the accuracy of a confidence interval. Suppose in an application you have $\ell_2 = 5$ excluded instruments and have estimated your equation by 2SLS. Now suppose that your reduced form F statistic equals 12. You check the Stock-Yogo table, and find that $F = 12$ is significant with $r = 0.20$. Thus we can interpret the conventional 2SLS confidence interval as having coverage of 80% (or 75% if we make the Bonferroni correction). On the other hand if $F = 27$ we would conclude that the test for weak instruments is significant with $r = 0.10$, meaning that the conventional 2SLS confidence interval can be interpreted as having coverage of 90% (or 85% after Bonferroni correction).

A more substantive extension, which we now discuss, reverses the steps. Unfortunately this discussion will be limited to the case $\ell_2 = 1$, where 2SLS and LIML are equivalent. First, use the reduced form F statistic to find a one-sided confidence interval for μ^2 of the form $[\mu_L^2, \infty)$. Second, use the lower bound μ_L^2 to calculate a critical value C for S_1 such that the 2SLS test has asymptotic size bounded below 0.05. This produces better size control than the Stock-Yogo procedure and produces more informative confidence intervals for β_2. We now describe the steps in detail.

The first goal is to find a one-sided confidence interval for μ^2. This is found by test inversion. As we described earlier, for any τ^2 we reject $\mathbb{H}_0 : \mu^2 = \tau^2$ in favor of $\mathbb{H}_1 : \mu^2 > \tau^2$ if $F > c$ where $G(c, \tau^2) = 0.95$. Equivalently, we reject if $G(F, \tau^2) > 0.95$. By the test inversion principle, an asymptotic 95% confidence interval $[\mu_L^2, \infty)$ can be formed as the set of all values of τ^2 which

are not rejected by this test. Since $G(F,\tau^2) \geq 0.95$ for all τ^2 in this set, the lower bound μ_L^2 satisfies $G(F,\mu_L^2) = 0.95$. The lower bound is found from this equation. Since this solution is not generally programmed, it needs to be found numerically. In MATLAB, the solution is `mu2` when `ncx2cdf(F,1,mu2)` returns `0.95`.

The second goal is to find the critical value C such that $\Pr(|S_1| \geq C) = 0.05$ when $\mu^2 = \mu_L^2$. From (11.79), this is achieved when

$$1 - G\left(\frac{\mu_L^2}{4} + C\mu_L, \frac{\mu_L^2}{4}\right) + G\left(\frac{\mu_L^2}{4} - C\mu_L, \frac{\mu_L^2}{4}\right) = 0.05. \tag{11.80}$$

This can be solved as

$$G\left(\frac{\mu_L^2}{4} + C\mu_L, \frac{\mu_L^2}{4}\right) = 0.95.$$

(The third term on the left-hand-side of (11.80) is zero for all solutions so can be ignored.) Using the non-central chi-square quantile function $Q(p,d)$, this C equals

$$C = \frac{Q\left(0.95, \frac{\mu_L^2}{4}\right) - \frac{\mu_L^2}{4}}{\mu_L}.$$

For example, in MATLAB this is found as `C=(ncx2inv(.95,1,mu2/4)-mu2/4)/sqrt(mu2)`. 95% confidence intervals for β_2 are then calculated as

$$\widehat{\beta}_{IV} \pm Cs(\widehat{\beta}_{\text{iv}}).$$

We can also calculate a p-value for the t-statistic T for β_2. These are

$$p = 1 - G\left(\frac{\mu_L^2}{4} + |T|\,\mu_L, \frac{\mu_L^2}{4}\right) + G\left(\frac{\mu_L^2}{4} - |T|\,\mu_L, \frac{\mu_L^2}{4}\right)$$

where the third term equals zero if $|T| \geq \mu_L/4$. In MATLAB, for example, this can be calculated by the commands

`T1 = mu2/4 + abs(T) * sqrt(mu2);`
`T2 = mu2/4 - abs(T) * sqrt(mu2);`
`p = -ncx2cdf(T1, 1, mu2/4) + ncx2cdf(T2, 1, mu2/4);`

These confidence intervals and p-values will be larger than the conventional intervals and p-values, reflecting the incorporation of information about the strength of the instruments through the first-stage F statistic. Also, by the Bonferroni bound these tests have asymptotic size bounded below 10% and the confidence intervals have asymptotic converage exceeding 90%, unlike the Stock-Yogo method which has size of 20% and coverage of 80%.

The augmented procedure suggested here, only for the $\ell_2 = 1$ case, is

1. Find μ_L^2 which solves $G\left(F, \mu_L^2\right) = 0.95$. In MATLAB, the solution is `mu2` when `ncx2cdf(F,1,mu2)` returns `0.95`.

2. Find C which solves $G\left(\mu_L^2/4 + C\mu_L, \mu_L^2/4\right) = 0.95$. In MATLAB, the command is
   `C=(ncx2inv(.95,1,mu2/4)-mu2/4)/sqrt(mu2)`

3. Report the confidence interval $\widehat{\beta}_2 \pm Cs(\widehat{\beta}_2)$ for β_2.

4. For the t statistic $T = \left(\widehat{\beta}_2 - \beta_2\right)/s(\widehat{\beta}_2)$ the asymptotic p-value is

   $$p = 1 - G\left(\frac{\mu_L^2}{4} + |T|\,\mu_L, \frac{\mu_L^2}{4}\right) + G\left(\frac{\mu_L^2}{4} - |T|\,\mu_L, \frac{\mu_L^2}{4}\right)$$

   which is computed in MATLAB by `T1=mu2/4+abs(T)*sqrt(mu2); T2=mu2/4-abs(T)*sqrt(mu2);` and `p=1-ncx2cdf(T1,1,mu2/4)+ncx2cdf(T2,1,mu2/4)`.

We have described an extension to the Stock-Yogo procedure for the case of one instrumental variable $\ell_2 = 1$. This restriction was due to the use of the analytic formula (11.80) for the asymptotic distribution, which is only available when $\ell_2 > 0$. In principle the procedure could be extended using simulation or bootstrap methods, but this has not been done to my knowledge.

To illustrate the Stock-Yogo and extended procedures, let us return to the Card proximity example. First, let's take the IV estimates reported in the second column of Table 11.1 which used *college* proximity as a single instrument. The reduced form estimates for the endogenous variable *education* is reported in the second column of Table 11.2. The excluded instrument *college* has a t-ratio of 4.2 which implies an F statistic of 17.8. The F statistic exceeds the rule-of thumb of 10, so the structural estimates pass the Stock-Yogo threshold. Based on the Stock-Yogo recommendation, this means that we can interpret the estimates conventionally. However, the conventional confidence interval, e.g. for the returns to education, $0.132 \pm 0.049 * 1.96 = [0.04, 0.23]$ has an asymptotic coverage of 80%, rather than the nominal 95% rate.

Now consider the extended procedure. Given $F = 17.8$ we can calculate the lower bound $\mu_L^2 = 6.6$. This implies a critical value of $C = 2.7$. Hence an improved confidence interval for the returns to education in this equation is $0.132 \pm 0.049 * 2.7 = [0.01, 0.26]$. This is a wider confidence interval, but has improved asymptotic coverage of 90%. The p-value for $\beta_2 = 0$ is $p = 0.012$.

Next, let's take the 2SLS estimates reported in the fourth column of Table 11.1 which use the two instruments *public* and *private*. The reduced form equation is reported in column six of Table 11.2. An F statistic for exclusion of the two instruments is $F = 13.9$, which exceeds the 15% size threshold for 2SLS and all thresholds for LIML, indicating that the structural estimates pass the Stock-Yogo threshold test and can be interpreted conventionally.

The weak instrument methods described here are important for applied econometrics as they discipline researchers to assess the quality of their reduced form relationships before reporting structural estimates. The theory, however, has limitations and shortcomings. A major limitation is that the theory requires the strong assumption of conditional homoskedasticity. Despite this theoretical limitation, in practice researchers apply the Stock-Yogo recommendations to estimates computed with heteroskedasticity-robust standard errors as it is the currently the best known approach. This is an active area of research so the recommended methods may change in the years ahead.

James Stock

James Stock (1955-) is a American econometrician and empirical macroeconomist who has made several important contributions, most notably his work on weak instruments, unit root testing, cointegration, and forecasting. He is also well-known for his undergraduate textbook *Introduction to Econometrics* (2014) co-authored with Mark Watson

11.32 Weak Instruments with $k_2 > 1$

When there are more than one endogenous regressor ($k_2 > 1$) it is better to examine the reduced form as a system. Staiger and Stock (1997) and Stock and Yogo (2005) provided an analysis of this case and constructed a test for weak instruments. The theory is considerably more involved than the $k_2 = 1$ case, so we briefly summarize it here excluding many details, emphasizing their suggested methods.

The structural equation and reduced form equations are

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i$$
$$\boldsymbol{x}_{2i} = \boldsymbol{\Gamma}_{12}'\boldsymbol{z}_{1i} + \boldsymbol{\Gamma}_{22}'\boldsymbol{z}_{2i} + \boldsymbol{u}_{2i}.$$

As in the previous section we assume that the errors are conditionally homoskedastic.

Identification of $\boldsymbol{\beta}_2$ requires the matrix $\boldsymbol{\Gamma}_{22}$ to be full rank. A necessary condition is that each row of $\boldsymbol{\Gamma}_{22}'$ is non-zero, but this is not sufficient.

We focus on the size performance of the homoskedastic Wald statistic for the 2SLS estimator of $\boldsymbol{\beta}_2$. For simplicity assume that the variance of e_i is known and normalized to one. Using representation (11.37), the Wald statistic can be written as

$$W = \boldsymbol{e}'\widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{X}_2\left(\boldsymbol{X}_2'\widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{e}\right)$$

where $\widetilde{\boldsymbol{Z}}_2 = (\boldsymbol{I}_n - \boldsymbol{P}_1)\boldsymbol{Z}_2$ and $\boldsymbol{P}_1 = \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$.

Stock and Staiger model the excluded instruments $\boldsymbol{z}_{2i}$ as weak by setting $\boldsymbol{\Gamma}_{22} = n^{-1/2}\boldsymbol{C}$ for some matrix $\boldsymbol{C}$. This is the multivariate analog of the simple case examined in the previous section. In this framework we have the asymptotic distribution results

$$\frac{1}{n}\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2 \xrightarrow{p} \boldsymbol{Q} = \mathbb{E}(\boldsymbol{z}_{2i}\boldsymbol{z}_{2i}') - \mathbb{E}(\boldsymbol{z}_{2i}\boldsymbol{z}_{1i}')\left(\mathbb{E}(\boldsymbol{z}_{1i}\boldsymbol{z}_{1i}')\right)^{-1}\mathbb{E}(\boldsymbol{z}_{1i}\boldsymbol{z}_{2i}')$$

$$\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{e} \xrightarrow{d} \boldsymbol{Q}^{1/2}\boldsymbol{\xi}_0$$

where $\boldsymbol{\xi}_0$ is a matrix normal variate whose columns are independent $\mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$. Furthermore, setting $\boldsymbol{\Sigma} = \mathbb{E}(\boldsymbol{u}_{2i}\boldsymbol{u}_{2i}')$ and $\overline{\boldsymbol{C}} = \boldsymbol{Q}^{1/2}\boldsymbol{C}\boldsymbol{\Sigma}^{-1/2}$,

$$\frac{1}{\sqrt{n}}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{X}_2 = \frac{1}{n}\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\,\boldsymbol{C} + \frac{1}{\sqrt{n}}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{U}_2 \xrightarrow{d} \boldsymbol{Q}^{1/2}\overline{\boldsymbol{C}}\boldsymbol{\Sigma}^{1/2} + \boldsymbol{Q}^{1/2}\boldsymbol{\xi}_2\boldsymbol{\Sigma}^{1/2}$$

where $\boldsymbol{\xi}_2$ is a matrix normal variates whose columns are independent $\mathrm{N}(\boldsymbol{0}, \boldsymbol{I})$. The variables $\boldsymbol{\xi}_0$ and $\boldsymbol{\xi}_2$ are correlated. Together we obtain the asymptotic distribution of the Wald statistic

$$W \xrightarrow{d} S = \boldsymbol{\xi}_0'\left(\overline{\boldsymbol{C}} + \boldsymbol{\xi}_2\right)\left(\overline{\boldsymbol{C}}'\overline{\boldsymbol{C}}\right)^{-1}\left(\overline{\boldsymbol{C}} + \boldsymbol{\xi}_2\right)'\boldsymbol{\xi}_0.$$

Using the spectral decomposition, $\overline{\boldsymbol{C}}'\overline{\boldsymbol{C}} = \boldsymbol{H}'\boldsymbol{\Lambda}\boldsymbol{H}$ where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ is diagonal. Thus we can write

$$S = \boldsymbol{\xi}_0'\overline{\boldsymbol{\xi}}_2\boldsymbol{\Lambda}^{-1}\overline{\boldsymbol{\xi}}_2'\boldsymbol{\xi}_0$$

where $\overline{\boldsymbol{\xi}}_2 = \overline{\boldsymbol{C}}\boldsymbol{H}' + \boldsymbol{\xi}_2\boldsymbol{H}'$. The matrix $\boldsymbol{\xi}^* = (\boldsymbol{\xi}_0, \overline{\boldsymbol{\xi}}_2)$ is multivariate normal, so $\boldsymbol{\xi}^{*\prime}\boldsymbol{\xi}^*$ has what is called a non-central Wishart distribution. It only depends on the matrix $\overline{\boldsymbol{C}}$ through $\boldsymbol{H}\overline{\boldsymbol{C}}'\overline{\boldsymbol{C}}\boldsymbol{H}' = \boldsymbol{\Lambda}$, which are the eigenvalues of $\overline{\boldsymbol{C}}'\overline{\boldsymbol{C}}$. Since S is a function of $\boldsymbol{\xi}^*$ only through $\overline{\boldsymbol{\xi}}_2'\boldsymbol{\xi}_0$ we conclude that S is a function of $\overline{\boldsymbol{C}}$ only through these eigenvalues.

This is a very quick derivation of a rather involved derivation, but the conclusion drawn by Stock and Yogo is that the asymptotic distribution of the Wald statistic is non-standard, and a function of the model parameters only through the eigenvalues of $\overline{\boldsymbol{C}}'\overline{\boldsymbol{C}}$ and the correlations between the normal variates $\boldsymbol{\xi}_0$ and $\overline{\boldsymbol{\xi}}_2$. The worst-case can be summarized by the maximal correlation between $\boldsymbol{\xi}_0$ and $\overline{\boldsymbol{\xi}}_2$ and the smallest eigenvalue of $\overline{\boldsymbol{C}}'\overline{\boldsymbol{C}}$. For convenience, they rescale the latter by dividing by the number of endogenous variables. Define

$$\boldsymbol{G} = \overline{\boldsymbol{C}}'\overline{\boldsymbol{C}}/k_2 = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{C}'\boldsymbol{Q}\boldsymbol{C}\boldsymbol{\Sigma}^{-1/2}/k_2$$

and

$$g = \lambda_{\min}(\boldsymbol{G}) = \lambda_{\min}\left(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{C}'\boldsymbol{Q}\boldsymbol{C}\boldsymbol{\Sigma}^{-1/2}\right)/k_2.$$

This can be estimated from the reduced-form regression

$$\boldsymbol{x}_{2i} = \widehat{\boldsymbol{\Gamma}}_{12}'\boldsymbol{z}_{1i} + \widehat{\boldsymbol{\Gamma}}_{22}'\boldsymbol{z}_{2i} + \widehat{\boldsymbol{u}}_{2i}.$$

The estimator is

$$\begin{aligned}
\widehat{\boldsymbol{G}} &= \widehat{\boldsymbol{\Sigma}}^{-1/2}\widehat{\boldsymbol{\Gamma}}_{22}'\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)\widehat{\boldsymbol{\Gamma}}_{22}\widehat{\boldsymbol{\Sigma}}^{-1/2}/k_2 \\
&= \widehat{\boldsymbol{\Sigma}}^{-1/2}\left(\boldsymbol{X}_2'\widetilde{\boldsymbol{Z}}_2\left(\widetilde{\boldsymbol{Z}}_2'\widetilde{\boldsymbol{Z}}_2\right)^{-1}\widetilde{\boldsymbol{Z}}_2'\boldsymbol{X}_2\right)\widehat{\boldsymbol{\Sigma}}^{-1/2}/k_2 \\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{n-k}\sum_{i=1}^{n}\widehat{\boldsymbol{u}}_{2i}\widehat{\boldsymbol{u}}_{2i}' \\
\widehat{g} &= \lambda_{\min}\left(\widehat{\boldsymbol{G}}\right).
\end{aligned}$$

$\widehat{\boldsymbol{G}}$ is a matrix F-type statistic for the coefficient matrix $\widehat{\boldsymbol{\Gamma}}_{22}$.

The statistic $\widehat{g}$ was proposed by Craig and Donald (1993) as a test for underidentification. Stock and Yogo (2005) use it as a test for weak instruments. Using simulation methods, they determined critical values for $\widehat{g}$ similar to those for the $k_2 = 1$ case. For given size $r > 0.05$, there is a critical value c (reported in the table below) such that if $\widehat{g} > c$, then the 2SLS (or LIML) Wald statistic W for $\widehat{\boldsymbol{\beta}}_2$ has asymptotic size bounded below r. On the other hand, if $\widehat{g} \leq c$ then we cannot bound the asymptotic size below r and we cannot reject the hypothesis of weak instruments.

The Stock-Yogo critical values for $k_2 = 2$ are presented in the following table. The methods and theory applies to the cases $k_2 > 2$ as well, but those critical values have not been calculated. As for the $k_2 = 1$ case, the critical values for 2SLS are dramatically increasing in ℓ_2. Thus when the model is over-identified, we need quite a large value of $\widehat{g}$ to reject the hypothesis of weak instruments. This is a strong cautionary message to check the $\widehat{g}$ statistic in applications. Furthermore, the critical values for LIML are generally decreasing in ℓ_2 (except for $r = 0.10$, where the critical values are increasing for large ℓ_2). This means that for over-identified models, LIML inference is much less sensitive to weak instruments than 2SLS, and may be the preferred estimation method.

The Stock-Yogo test can be implemented in Stata for $k_2 \leq 2$ using the command `estat firststage` after `ivregress 2sls` or `ivregres liml` if a standard (non-robust) covariance matrix has been specified (that is, without the '`,r`' option).

$\widehat{g}$ 5% Critical Value for Weak Instruments, $k_2 = 2$

	Maximal Size r							
	2SLS				LIML			
ℓ_2	0.10	0.15	0.20	0.25	0.10	0.15	0.20	0.25
2	7.0	4.6	3.9	3.6	7.0	4.6	3.9	3.6
3	13.4	8.2	6.4	5.4	5.4	3.8	3.3	3.1
4	16.9	9.9	7.5	6.3	4.7	3.4	3.0	2.8
5	19.4	11.2	8.4	6.9	4.3	3.1	2.8	2.6
6	21.7	12.3	9.1	7.4	4.1	2.9	2.6	2.5
7	23.7	13.3	9.8	7.9	3.9	2.8	2.5	2.4
8	25.6	14.3	10.4	8.4	3.8	2.7	2.4	2.3
9	27.5	15.2	11.0	8.8	3.7	2.7	2.4	2.2
10	29.3	16.2	11.6	9.3	3.6	2.6	2.3	2.1
15	38.0	20.6	14.6	11.6	3.5	2.4	2.1	2.0
20	46.6	25.0	17.6	13.8	3.6	2.4	2.0	1.9
25	55.1	29.3	20.6	16.1	3.6	2.4	1.97	1.8
30	63.5	33.6	23.5	18.3	4.1	2.4	1.95	1.7

11.33 Many Instruments

Some applications have available a large number ℓ of instruments. If they are all valid, using a large number should reduce the asymptotic variance relative to estimation with a smaller number of instruments. Is it then good practice to use many instruments? Or is there a cost to this practice? Bekker (1994) initiated a large literature investigating this question by formalizing the idea of "many instruments". Bekker proposed an asymptotic approximation which treats the number of instruments ℓ as proportional to the sample size, that is $\ell = \alpha n$, or equivalently that $\ell/N \to \alpha \in [0, 1)$.

We examine this idea in the simplified setting of one endogenous regressor and no included exogenous regressors

$$\begin{aligned} y_i &= \beta x_i + e_i \\ x_i &= \boldsymbol{z}_i'\boldsymbol{\gamma} + u_i \end{aligned} \tag{11.81}$$

with $\boldsymbol{z}_i$ $\ell \times 1$. As in the previous two sections we make the simplifying assumption that the errors are conditionally homoskedastic and unit variance

$$\operatorname{var}\left(\begin{pmatrix} e_i \\ u_i \end{pmatrix} \mid \boldsymbol{z}_i\right) = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{11.82}$$

In addition we assume that the conditional fourth moments are bounded

$$\mathbb{E}\left(e_i^4 \mid \boldsymbol{z}_i\right) \leq C < \infty \qquad \mathbb{E}\left(u_i^4 \mid \boldsymbol{z}_i\right) \leq C < \infty. \tag{11.83}$$

The idea that there are "many instruments" is formalized by the assumption that the number of instruments is increasing proportionately with the sample size

$$\frac{\ell}{n} \longrightarrow \alpha. \tag{11.84}$$

The best way to think about this is to view α as the ratio of ℓ to n in a given sample. Thus if an application has $n = 100$ observations and $\ell = 10$ instruments, then we should treat $\alpha = 0.10$.

Consider the variance of the endogenous regressor x_i from the reduced form: $\operatorname{var}(x_i) = \operatorname{var}(\boldsymbol{z}_i'\boldsymbol{\gamma}) + \operatorname{var}(u_i)$. Suppose that $\operatorname{var}(x_i)$ and $\operatorname{var}(u_i)$ are unchanging as ℓ increases. This implies that $\operatorname{var}(\boldsymbol{z}_i'\boldsymbol{\gamma})$ is unchanging as well. This will be a useful assumption, as it implies that the population R^2 of the reduced form is not changing with ℓ. We don't need this exact condition, rather we simply assume that the sample version converges in probability to a fixed constant

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\gamma}'\boldsymbol{z}_i\boldsymbol{z}_i'\boldsymbol{\gamma} \xrightarrow{p} c \tag{11.85}$$

for $0 < c < \infty$. Again, this essentially implies that the R^2 of the reduced form regression for x_i converges to a constant.

As a baseline it is useful to examine the behavior of the least-squares estimator of β. First, observe that the variances of $n^{-1}\sum_{i=1}^{n} \boldsymbol{\gamma}'\boldsymbol{z}_i e_i$ and $n^{-1}\sum_{i=1}^{n} \boldsymbol{\gamma}'\boldsymbol{z}_i u_i$, conditional on $\boldsymbol{Z}$, are both equal to

$$n^{-2}\sum_{i=1}^{n} \boldsymbol{\gamma}'\boldsymbol{z}_i\boldsymbol{z}_i'\boldsymbol{\gamma} \xrightarrow{p} 0$$

by (11.85). Thus they converge in probability to zero:

$$n^{-1}\sum_{i=1}^{n} \boldsymbol{\gamma}'\boldsymbol{z}_i e_i \xrightarrow{p} 0 \tag{11.86}$$

and

$$n^{-1}\sum_{i=1}^{n}\boldsymbol{\gamma}'\boldsymbol{z_i}\boldsymbol{u}_i \xrightarrow{p} 0. \tag{11.87}$$

Combined with (11.85) and the WLLN we find

$$\frac{1}{n}\sum_{i=1}^{n}x_ie_i = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}'\boldsymbol{z_i}e_i + \frac{1}{n}\sum_{i=1}^{n}u_ie_i \xrightarrow{p} \rho$$

$$\frac{1}{n}\sum_{i=1}^{n}x_i^2 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}'\boldsymbol{z_i}\boldsymbol{z}_i'\boldsymbol{\gamma} + \frac{2}{n}\sum_{i=1}^{n}\boldsymbol{\gamma}'\boldsymbol{z_i}u_i + \frac{1}{n}\sum_{i=1}^{n}u_i^2 \xrightarrow{p} c+1.$$

Hence

$$\widehat{\beta}_{\text{ols}} = \beta + \frac{\frac{1}{n}\sum_{i=1}^{n}x_ie_i}{\frac{1}{n}\sum_{i=1}^{n}x_i^2} \xrightarrow{p} \beta + \frac{\rho}{c+1}.$$

Thus least-squares is inconsistent for β under endogeneity.

Now consider the 2SLS estimator. In matrix notation, setting $\boldsymbol{P} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$,

$$\widehat{\beta}_{\text{2sls}} - \beta = \frac{\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{e}}{\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}} = \frac{\frac{1}{n}\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{e} + \frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{e}}{\frac{1}{n}\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{\gamma} + \frac{2}{n}\boldsymbol{\gamma}'\boldsymbol{Z}'\boldsymbol{u} + \frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}}. \tag{11.88}$$

In the expression on the right-side of (11.88), three of the components have been examined in (11.85), (11.86), and (11.87). We now examine the remaining components $\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{e}$ and $\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{e} = \boldsymbol{u}$.

First, it it simple to take their expectations under the conditional homoskedasticity assumption. We have

$$\mathbb{E}\left(\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{e}\right) = \frac{1}{n}\operatorname{tr}\mathbb{E}\left(\boldsymbol{P}\boldsymbol{e}\boldsymbol{u}'\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{P}\right)\rho = \frac{\ell}{n}\rho \tag{11.89}$$

since $\operatorname{tr}\left(\boldsymbol{P}\right) = \ell$. Similarly

$$\mathbb{E}\left(\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}\right) = \frac{1}{n}\operatorname{tr}\mathbb{E}\left(\boldsymbol{P}\boldsymbol{u}\boldsymbol{u}'\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{P}\right) = \frac{\ell}{n}.$$

Second, we examine their variances, which is a more cumbersome exercise. Let $P_{ij} = \boldsymbol{z}_i'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_j$ be the ij^{th} element of $\boldsymbol{P}$. Then $\boldsymbol{u}'\boldsymbol{P}\boldsymbol{e} = \sum_{i=1}^{n}\sum_{j=1}^{n}u_ie_jP_{ij}$ and $\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u} = \sum_{i=1}^{n}\sum_{j=1}^{n}u_iu_jP_{ij}$.

The matrix $\boldsymbol{P}$ is idempotent. It therefore has the properties $\sum_{i=1}^{n}P_{ii} = \operatorname{tr}\left(\boldsymbol{P}\right) = \ell$ and $0 \le P_{ii} \le 1$. The property $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$ also implies $\sum_{j=1}^{n}P_{ij}^2 = P_{ii}$. Then

$$\begin{aligned}
\operatorname{var}\left(\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{e}\right) &= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\left(u_ie_j - \rho 1\left(i=j\right)\right)P_{ij}\right)^2 \\
&= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n}\left(u_ie_j - \rho 1\left(i=j\right)\right)P_{ij}\left(u_ke_l - \rho 1\left(k=l\right)\right)P_{kl}\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(\left(u_ie_i - \rho\right)^2P_{ii}^2\right) && (11.90) \\
&+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\left(u_i^2e_j^2P_{ij}^2\right) && (11.91) \\
&+ \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\left(u_ie_je_iu_jP_{ij}^2\right) && (11.92)
\end{aligned}$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(u_i^2 e_i^2 P_{ii}^2\right) - 2\frac{\rho}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(u_i e_i P_{ii}^2\right) + \rho^2\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(P_{ii}^2\right).$$

The third equality holds because the remaining cross-products have zero expectation since the observations are independent and the errors have zero mean. We then calculate that (11.90) is bounded by

$$\left(C-\rho^2\right)\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}P_{ii}^2 \le \left(C-\rho^2\right)\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(P_{ii}\right) = \left(C-\rho^2\right)\frac{\ell}{n^2} \longrightarrow 0$$

under (11.84). The first inequality is $P_{ii} \le 1$ and the equality is $\sum_{i=1}^{n} P_{ii} = \ell$. Next, the conditional homoskedasticity assumption implies that (11.91) plus (11.92) equals $\left(1+\rho^2\right)$ times

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}\mathbb{E}\left(P_{ij}^2\right) \le \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left(P_{ij}^2\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left(P_{ii}\right) = \frac{\ell}{n^2} \longrightarrow 0$$

under (11.84). The first equality is $\sum_{j=1}^{n} P_{ij}^2 = P_{ii}$. Together, we have shown that

$$\operatorname{var}\left(\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{e}\right) \longrightarrow 0.$$

Using (11.89) and Markov's inequality

$$\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{e} - \frac{\ell}{n}\rho \xrightarrow{p} 0.$$

Combined with (11.84) we find

$$\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{e} \xrightarrow{p} \alpha\rho. \tag{11.93}$$

The analysis for $\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u}$ is quite similar. We deduce that

$$\frac{1}{n}\boldsymbol{u}'\boldsymbol{P}\boldsymbol{u} \xrightarrow{p} \alpha. \tag{11.94}$$

Returning to the 2SLS estimator (11.88) and combining (11.85), (11.86), (11.87), (11.93) and (11.94), we find

$$\widehat{\beta}_{2\text{sls}} \xrightarrow{p} \beta + \frac{\alpha\rho}{c+\alpha}.$$

We can state this formally.

Theorem 11.33.1 *In model (11.81), under assumptions (11.82), (11.83) and (11.84), then as $n \to \infty$.*

$$\widehat{\beta}_{\text{ols}} \xrightarrow{p} \beta + \frac{\rho}{c+1}$$

$$\widehat{\beta}_{2\text{sls}} \xrightarrow{p} \beta + \frac{\alpha\rho}{c+\alpha}.$$

This result is quite insightful. It shows that while endogeneity ($\rho \neq 0$) renders the least-squares estimator inconsistent, the 2SLS estimator is also inconsistent if the number of instruments diverges proportionately with n. The limit in Theorem 11.33.1 shows a continuity between least-squares and 2SLS. The probability limit of the 2SLS estimator is continuous in α, with the extreme case ($\alpha = 1$)

implying that 2SLS and least-squares have the same probability limit. The general implication is that the inconsistency of 2SLS is increasing in α.

Hence using a large number of instruments in an application comes at a cost.

In an application, users should calculate the "many instrument ratio" $\alpha = \ell/n$. Unfortunately there is no known rule-of-thumb for α which should lead to acceptable inference, but a minimum criterion is that if $\alpha \geq 0.05$ you should be seriously concerned about the many-instrument problem. In general, if it is desired to use a large number of instruments then it is recommended to use an estimation method other than 2SLS such as LIML.

11.34 Example: Acemoglu, Johnson and Robinson (2001)

One particularly well-cited instrument variable regression is in Acemoglu, Johnson and Robinson (2001) with additional details published in (2012). They are interested in the effect of political institutions on economic performance. The theory is that good institutions (rule-of-law, property rights) should result in a country having higher long-term economic output than if the same country had poor institutions. To investigate this question, they focus on a sample of 64 former European colonies. Their data is in the file `AJR2001` on the textbook website.

The authors' premise is that modern political institutions will have been influenced by the colonizing country. In particular, they argue that colonizing countries tended to set up colonies as either an "extractive state" or as a "migrant colony". An extractive state was used by the colonizer to extract resources for the colonizing country, but was not largely settled by the European colonists. In this case the colonists would have had no incentive to set up good political institutions. In contrast, if a colony was set up as a "migrant colony", then large numbers of European settlers migrated to the colony to live. These settlers would have desired institutions similar to those in their home country, and hence would have had a positive incentive to set up good political institutions. The nature of institutions is quite persistent over time, so these 19^{th}-century foundations would affect the nature of modern institutions. The authors conclude that the 19^{th}-century nature of the colony should be predictive of the nature of modern institutions, and hence modern economic growth.

To start the investigation they report an OLS regression of log GDP per capita in 1995 on a measure of political institutions they call "risk", which is a measure of the protection against expropriation risk. This variable ranges from 0 to 10, with 0 the lowest protection against appropriation, and 10 the highest. For each country the authors take the average value of the index over 1985 to 1995 (the mean is 6.5 with a standard deviation of 1.5). Their reported OLS estimates (intercept omitted) are

$$\widehat{\log(GDP\ per\ Capita)} = \underset{(0.06)}{0.52}\ risk. \tag{11.95}$$

These estimates imply a 52% difference in GDP between countries with a 1-unit difference in *risk*.

The authors argue that the *risk* is likely endogenous, since economic output influences political institutions, and because the variable *risk* is undoubtedly measured with error. These issues induce least-square bias in different directions and thus the overall bias effect is unclear.

To correct for the endogeneity bias the authors argue the need for an instrumental variable which does not directly affect economic performance yet is associated with political institutions. Their innovative suggestion was to use the mortality rate which faced potential European settlers in the 19^{th} century. Colonies with high expected mortality would have been less attractive to European setters, resulting in lower levels of European migrants. As a consequence the authors expect such colonies to have been more likely structured as an extractive state rather than a migrant colony. To measure the expected mortality rate the authors use estimates provided by historical research of the annualized deaths per 1000 soldiers, labeled *mortality*. (They used military mortality rates

as the military maintained high-quality records.) The first-stage regression is

$$risk = \underset{(0.13)}{-0.61} \log(mortality) + \widehat{u}. \tag{11.96}$$

These estimates confirm that 19^{th}-century high settler mortality rates are associated with countries with lower quality modern institutions. Using $\log(mortality)$ as an instrument for $risk$, they estimate the structural equation using 2SLS and report

$$\widehat{\log(GDP\ per\ Capita)} = \underset{(0.16)}{0.94}\ risk. \tag{11.97}$$

This estimate is much higher than the OLS estimate from (11.95). The estimate is consistent with a near doubling of GDP due to a 1-unit difference in the risk index.

These are simple regressions involving just one right-hand-side variable. The authors considered a range of other models. Included in these results are a reversal of a traditional finding. In a conventional (least-squares) regression two relevant varibles for output are *latitude* (distance from the equator) and *africa* (a dummy variable for countries from Africa), both of which are difficult to interpret causally. But in the proposed instrumental variables regression the variables *latitude* and *africa* have much smaller – and statistically insignificant – coefficients.

To assess the specification, we can use the Stock-Yogo and endogeneity tests. The Stock-Yogo test is from the reduced form (11.96). The instrument has a t-ratio of 4.8 (or $F = 23$) which exceeds the Stock-Yogo critical value and hence can be treated as strong. For an endogeneity test, we take the least-squares residual $\widehat{u}$ from this equation and include it in the structural equation and estimate by least-squares. We find a coefficient on $\widehat{u}$ of -0.57 with a t-ratio of 4.7, which is highly significant. We conclude that the least-squares and 2SLS estimates are statistically different, and reject the hypothesis that the variable *risk* is exogenous for the GDP structural equation.

In Exercise 11.23 you will replicate and extend these results using the authors' data.

This paper is a creative and careful use of the instrumental variables method. The creativity stems from the careful historical analysis which lead to the focus on mortality as a potential predictor of migration choices. The care comes in the implementation, as the authors needed to gather country-level data on political institutions and mortality from distinct sources. Putting these pieces together is the art of the project.

11.35 Example: Angrist and Krueger (1991)

Another influential instrument variable regression is in Angrist and Krueger (1991). Their concern, similar to Card (1995), is estimation of the structural returns to education while treating educational attainment as endogenous. Like Card, their goal is to find an instrument which is exogenous for wages yet has an impact on educational attainment. A subset of their data in the file `AK1991` on the textbook website.

Their creative suggestion was to focus on compulsory school attendance policies and their interaction with birthdates. Compulsory schooling laws vary across states in the United States, but typically require that youth remain in school until their sixteenth or seventeenth birthday. Angrist and Krueger argue that compulsory schooling has a causal effect on wages – youth who would have chosen to drop out of school stay in school for more years – and thus have more education which causally impacts their earnings as adults.

Angrist and Krueger next observe that these policies have differential impact on youth who are born early or late in the school year. Students who are born early in the calendar year are typically older when they enter school. Conseqeuntly when they attain the legal dropout age they

have attended less school than those born near the end of the year. This means that birthdate (early in the calendar year versus late) exogenously impacts educational attainment, and thus wages through education. Yet birthdate must be exogenous for the structural wage equation, as there is no reason to believe that birthdate itself has a causal impact on a person's ability or wages. These considerations together suggest that birthdate is a valid instrumental variable for education in a causal wage equation.

Typical wage datasets include age, but not birthdates. To obtain information on birthdate, Angrist and Krueger used a U.S. Census data which includes an individual's quarter of birth (January-March, April-June, etc.). They use this variable to construct 2SLS estimates of the return to education.

Their paper carefully documents that educational attainment varies by quarter of birth (as predicted by the above discussion), and reports a large set of least-squares and 2SLS estimates. We focus on two estimates at the core of their analysis, reported in column (6) of their Tables V and VII. This involves data from the 1980 census with men born in 1930-1939, with 329,509 observations. The first equation is

$$\begin{array}{rccccccccc} \widehat{\log(wage)} = & 0.080 & edu - & 0.230 & black + & 0.158 & urban + & 0.244 & married \\ & (0.016) & & (0.026) & & (0.017) & & (0.005) & \end{array} \tag{11.98}$$

where *edu* years of education, and *black*, *urban*, and *married* are dummy variables indicating race (1 if black, 0 otherwise), lives in a metropolitan area, and if married. In addition to the reported coefficients, the equation also includes as regressors nine year-of-birth dummies and eight region-of-residence dummies. The equation is estimated by 2SLS. The instrumental variables are the 30 interactions of three quarter-of-birth times ten year-of-birth dummy variables.

This equation indicates an 8% increase in wages due to each year of education.

Angrist and Krueger observe that the effect of compulsory education laws are likely to vary across states, so expand the instrument set to include interactions with state-of-birth. They estimate the following equation by 2SLS

$$\begin{array}{rccccccccc} \widehat{\log(wage)} = & 0.083 & edu - & 0.233 & black + & 0.151 & urban + & 0.244 & married. \\ & (0.010) & & (0.011) & & (0.010) & & (0.003) & \end{array} \tag{11.99}$$

This equation also adds fifty state-of-birth dummy variables as regressors. The instrumental variables are the 180 interactions of quarter-of-birth times year-of-birth dummy variables, plus quarter-of-birth times state-of-birth interactions.

This equation shows a similar estimated causal effect of education on wages as in (11.98). More notably, the standard error is smaller in (11.99), suggesting improved precision by the expanded instrumental variable set.

However, these estimates seem excellent candidates for weak instruments and many instruments. Indeed, this paper (published in 1991) helped sparked these two literatures. We can use the Stock-Yogo tools to explore the instrument strength and the implications for the Angrist-Krueger estimates.

We first take equation (11.98). Using the original Angrist-Krueger data, we estimate the corresponding reduced form, and calculate the F statistic for the 30 excluded instruments. We find $F = 4.7$. It has an asymptotic p-value of 0.000, suggesting that we can reject (at any significance level) the hypothesis that the coefficients on the excluded instruments are zero. Thus Angrist and Krueger appear to be correct that quarter of birth helps to explain educational attainment and are thus a valid instrumental variable set. However, using the Stock-Yogo test, $F = 4.7$ is not high enough to reject the hypothesis that the instruments are weak. Specifically, for $\ell_2 = 30$ the critical value for the F statistic is 45 (if we want to bound size below 15%). The actual value of 4.7 is

far below 45. Since we cannot reject that the instruments are weak, this indicates that we cannot interpret the 2SLS estimates and test statistics in (11.98) as reliable.

Second, take (11.99) with the expanded regressor and instrument set. Estimating the corresponding reduced form, we find the F statistic for the 180 excluded instruments is $F = 2.15$ which also has an asymptotic p-value of 0.000 indicating that we can reject at any significance level the hypothesis that the excluded instruments have no effect on educational attainment. However, using the Stock-Yogo test we also cannot reject the hypothesis that the instruments are weak. While Stock and Yogo did not calculate the critical values for $\ell_2 = 180$, the 2SLS critical values are increasing in ℓ_2 so we we can use those for $\ell_2 = 30$ as a lower bound. Hence the observed value of $F = 2.15$ is far below the level needed for significance. Consequently the results in (11.99) cannot be viewed as reliable. In particular, the observation that the standard errors in (11.99) are smaller than those in (11.98) should not be interpreted as evidence of greater precision. Rather, they should be viewed as evidence of unreliability due to weak instruments.

When instruments are weak, one constructive suggestion is to use LIML estimation rather than 2SLS. Another constructive suggestion is to alter the instrument set. While Angrist and Krueger used a large number of instrumental variables, we can consider using a smaller set. Take equation (11.98). Rather than estimating it using the 30 interaction instruments, consider using only the three quarter-of-birth dummy variables. We report the reduced form estimates here:

$$\begin{array}{cccccccccccc} \widehat{edu} = - & 1.57 & black + & 1.05 & urban + & 0.225 & married + & 0.050 & Q_2 + & 0.101 & Q_3 + & 0.142 \quad Q_4 \\ & (0.02) & & (0.01) & & (0.016) & & (0.016) & & (0.016) & & (0.016) \end{array} \tag{11.100}$$

where Q_2, Q_3 and Q_4 are dummy variables for birth in the 2^{nd}, 3^{rd}, and 4^{th} quarter. The regression also includes nine year-of-birth and eight region-of-residence dummy variables.

The reduced form coefficients in (11.100) on the quarter-of-birth dummies are quite instructive. The coefficients are positive and increasing, consistent with the Angrist-Krueger hypothesis that individuals born later in the year achieve higher average education. Focusing on the weak instrument problem, the F test for exclusion of these three variables is $F = 30$. The Stock-Yogo critical value is 12.8 for $\ell_2 = 3$ and a size of 15%, and is 22.3 for a size of 10%. Since $F = 30$ exceeds both these thresholds we can reject the hypothesis that this reduced form is weak. Estimating the model by 2SLS with these three instruments we find

$$\begin{array}{cccccccccc} \widehat{\log(wage)} = & 0.098 & edu - & 0.217 & black + & 0.137 & urban + & 0.240 & married. \\ & (0.020) & & (0.022) & & (0.017) & & (0.006) & \end{array} \tag{11.101}$$

These estimates indicate a slightly larger (10%) causal impact of education on wages, but with a larger standard error. The Stock-Yogo analysis indicates that we can interpret the confidence intervals from these estimates as having asymptotic coverge 85%.

While the original Angrist-Krueger estimates suffer due to weak instruments, their paper is a very creative and thoughtful application of the **natural experiment** methodology. They discovered a completely exogenous variation present in the world – birthdate – and showed how this has a small but measurable effect on educational attainment, and thereby on earnings. Their crafting of this natural experiment regression is extremely clever and demonstrates a style of analysis which can successfully underlie an effective instrumental variables empirical analysis.

Joshua Angrist

Joshua Angrist (1960-) is an Israeli-American econometrician and labor economist who is known for his advocacy of natural experiments to motivate instrumental variables estimation. He is also well-known for his book *Mostly Harmless Econometrics* (2009) co-authored with Jörn-Steffen Pischke.

11.36 Programming

We now present Stata code for some of the empirical work reported in this chapter.

Stata do File for Card Example

```
use Card1995.dta, clear
set more off
gen exp = age76 - ed76 - 6
gen exp2 = (exp^2)/100
* Drop observations with missing wage
drop if lwage76==.
* Least squares baseline
reg lwage76 ed76 exp exp2 smsa76r reg76r, r
* Reduced form estimates using college as instrument
reg lwage76 nearc4 exp exp2 smsa76r reg76r, r
reg ed76 nearc4 exp exp2 smsa76r reg76r, r
* IV estimates
ivregress 2sls lwage76 exp exp2 smsa76r reg76r (ed76=nearc4), r
* Reduced form using public and private as instruments
reg ed76 nearc4a nearc4b exp exp2 smsa76r reg76r, r
* F test for excluded instruments
testparm nearc4a nearc4b
predict u2, residual
* 2SLS estimates using both instruments
ivregress 2sls lwage76 exp exp2 smsa76r reg76r (ed76=nearc4a nearc4b), r
* Control function regressions
reg lwage76 ed76 exp exp2 smsa76r reg76r u2
reg lwage76 ed76 exp exp2 smsa76r reg76r u2, r
* LIML estimates
ivregress liml lwage76 exp exp2 smsa76r reg76r (ed76=nearc4a nearc4b), r
```

Stata do File for Acemoglu-Johnson-Robinson Example

```
use AJR2001.dta, clear
reg loggdp risk
reg risk logmort0
predict u, residual
ivregress 2sls loggdp (risk=logmort0)
reg loggdp risk u
```

Stata do File for Angrist-Krueger Example

```
use AK1991.dta, clear
ivregress 2sls logwage black smsa married i.yob i.region (edu = i.qob#i.yob)
reg edu black smsa married i.yob i.region i.qob#i.yob
testparm i.qob#i.yob
ivregress 2sls logwage black smsa married i.yob i.region i.state (edu = i.qob#i.yob i.qob#i.state)
reg edu black smsa married i.yob i.region i.state i.qob#i.yob i.qob#i.state
testparm i.qob#i.yob i.qob#i.state
reg edu black smsa married i.yob i.region i.qob
testparm i.qob
ivregress 2sls logwage black smsa married i.yob i.region (edu = i.qob)
```

Exercises

Exercise 11.1 Consider the single equation model

$$y_i = z_i\beta + e_i,$$

where y_i and z_i are both real-valued (1×1). Let $\widehat{\beta}$ denote the IV estimator of β using as an instrument a dummy variable d_i (takes only the values 0 and 1). Find a simple expression for the IV estimator in this context.

Exercise 11.2 In the linear model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \end{aligned}$$

suppose $\sigma_i^2 = \mathbb{E}\left(e_i^2 \mid x_i\right)$ is known. Show that the GLS estimator of $\boldsymbol{\beta}$ can be written as an IV estimator using some instrument $\boldsymbol{z}_i$. (Find an expression for $\boldsymbol{z}_i$.)

Exercise 11.3 Take the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}.$$

Let the OLS estimator for $\boldsymbol{\beta}$ be $\widehat{\beta}$ and the OLS residual be $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

Let the IV estimator for $\boldsymbol{\beta}$ using some instrument $\boldsymbol{Z}$ be $\widetilde{\boldsymbol{\beta}}$ and the IV residual be $\widetilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$. If $\boldsymbol{X}$ is indeed endogenous, will IV "fit" better than OLS, in the sense that $\widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}} < \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$, at least in large samples?

Exercise 11.4 The reduced form between the regressors $\boldsymbol{x}_i$ and instruments $\boldsymbol{z}_i$ takes the form

$$\boldsymbol{x}_i = \boldsymbol{\Gamma}'\boldsymbol{z}_i + \boldsymbol{u}_i$$

or

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U}$$

where $\boldsymbol{x}_i$ is $k \times 1$, $\boldsymbol{z}_i$ is $l \times 1$, $\boldsymbol{X}$ is $n \times k$, $\boldsymbol{Z}$ is $n \times l$, $\boldsymbol{U}$ is $n \times k$, and $\boldsymbol{\Gamma}$ is $l \times k$. The parameter $\boldsymbol{\Gamma}$ is defined by the population moment condition

$$\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{u}_i'\right) = \boldsymbol{0}$$

Show that the method of moments estimator for $\boldsymbol{\Gamma}$ is $\widehat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$.

Exercise 11.5 In the structural model

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \boldsymbol{X} &= \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U} \end{aligned}$$

with $\boldsymbol{\Gamma}$ $l \times k$, $l \geq k$, we claim that $\boldsymbol{\beta}$ is identified (can be recovered from the reduced form) if $\operatorname{rank}(\boldsymbol{\Gamma}) = k$. Explain why this is true. That is, show that if $\operatorname{rank}(\boldsymbol{\Gamma}) < k$ then $\boldsymbol{\beta}$ cannot be identified.

Exercise 11.6 For Theorem 11.16.1, establish that $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$.

Exercise 11.7 Take the linear model

$$\begin{aligned} y_i &= x_i\beta + e_i \\ \mathbb{E}\left(e_i \mid x_i\right) &= 0. \end{aligned}$$

where x_i and β are 1×1.

(a) Show that $\mathbb{E}(x_i e_i) = 0$ and $\mathbb{E}\left(x_i^2 e_i\right) = 0$. Is $\boldsymbol{z}_i = (x_i \quad x_i^2)'$ a valid instrumental variable for estimation of β?

(b) Define the 2SLS estimator of β, using $\boldsymbol{z}_i$ as an instrument for x_i. How does this differ from OLS?

Exercise 11.8 Suppose that price and quantity are determined by the intersection of the linear demand and supply curves

$$\begin{aligned} \text{Demand}: &\quad Q = a_0 + a_1 P + a_2 Y + \boldsymbol{e}_1 \\ \text{Supply}: &\quad Q = b_0 + b_1 P + b_2 W + \boldsymbol{e}_2 \end{aligned}$$

where income (Y) and wage (W) are determined outside the market. In this model, are the parameters identified?

Exercise 11.9 Consider the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(e_i|\boldsymbol{z}_i) &= 0 \end{aligned}$$

with y_i scalar and $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ each a k vector. You have a random sample $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i : i = 1, ..., n)$.

(a) Suppose that $\boldsymbol{x}_i$ is exogeneous in the sense that $E(e_i|\boldsymbol{z}_i, \boldsymbol{x}_i) = 0$. Is the IV estimator $\widehat{\boldsymbol{\beta}}_{\text{iv}}$ unbiased for $\boldsymbol{\beta}$?

(b) Continuing to assume that $\boldsymbol{x}_i$ is exogeneous, find the variance matrix for $\widehat{\boldsymbol{\beta}}_{\text{iv}}$, var $\left(\widehat{\boldsymbol{\beta}}_{\text{iv}}|\boldsymbol{X}, \boldsymbol{Z}\right)$.

Exercise 11.10 Consider the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \boldsymbol{x}_i &= \boldsymbol{\Gamma}'\boldsymbol{z}_i + \boldsymbol{u}_i \\ \mathbb{E}(\boldsymbol{z}_i e_i) &= \boldsymbol{0} \\ \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{u}_i'\right) &= \boldsymbol{0} \end{aligned}$$

with y_i scalar and $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ each a k vector. You have a random sample $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i : i = 1, ..., n)$. Take the control function equation

$$\begin{aligned} e_i &= \boldsymbol{u}_i'\boldsymbol{\gamma} + \varepsilon_i \\ \mathbb{E}(\boldsymbol{u}_i \varepsilon_i) &= \boldsymbol{0} \end{aligned}$$

and assume for simplicity that $\boldsymbol{u}_i$ is observed. Inserting into the structural equation we find

$$y_i = \boldsymbol{z}_i'\boldsymbol{\beta} + \boldsymbol{u}_i'\boldsymbol{\gamma} + \varepsilon_i$$

The control function estimator $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$ is OLS estimation of this equation.

(a) Show that $\mathbb{E}(\boldsymbol{x}_i \varepsilon_i) = \boldsymbol{0}$ (algebraically)

(b) Derive the asymptotic distribution of $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}})$.

Exercise 11.11 Consider the structural equation

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \tag{11.102}$$

with x_i treated as endogenous so that $E(x_i e_i) \neq 0$. Assume y_i and x_i are scalar. Suppose we also have a scalar instructment z_i which satisfies

$$\mathbb{E}(e_i|z_i) = 0$$

so in particular $\mathbb{E}(e_i) = 0$, $\mathbb{E}(z_i e_i) = 0$ and $\mathbb{E}\left(z_i^2 e_i\right) = 0$.

(a) Should x_i^2 be treated as endogenous or exogenous?

(b) Suppose we have a scalar instrument z_i which satisfies

$$x_i = \gamma_0 + \gamma_1 z_i + u_i \tag{11.103}$$

with u_i independent of z_i and mean zero.

Consider using $(1, z_i, z_i^2)$ as instruments. Is this a sufficient number of instruments? (Would this be just-identified, over-identified, or under-identified)?

(c) Write out the reduced form equation for x_i^2. Under what condition on the reduced form parameters (11.103) are the parameters in (11.102) identified?

Exercise 11.12 Consider the structural equation and reduced form

$$\begin{aligned} y_i &= \beta x_i^2 + e_i \\ x_i &= \gamma z_i + u_i \\ \mathbb{E}\left(z_i e_i\right) &= 0 \\ \mathbb{E}\left(z_i u_i\right) &= 0 \end{aligned}$$

with x_i^2 treated as endogenous so that $\mathbb{E}\left(x_i^2 e_i\right) \neq 0$. For simplicity assume no intercepts. y_i, z_i, and x_i are scalar. Assume $\gamma \neq 0$. Consider the following estimator. First, estimate γ by OLS of x_i on z_i and construct the fitted values $\widehat{x}_i = \widehat{\gamma} z_i$. Second, estimate β by OLS of y_i on $\widehat{x}_i^2$.

(a) Write out this estimator $\widehat{\beta}$ explicitly as a function of the sample

(b) Find its probability limit as $n \to \infty$

(c) In general, is $\widehat{\beta}$ consistent for β? Is there a reasonable condition under which $\widehat{\beta}$ is consistent?

Exercise 11.13 Consider the structural equation

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= 0 \end{aligned}$$

where $\boldsymbol{x}_{2i}$ is $k_2 \times 1$ and treated as endogenous. The variables $\boldsymbol{z}_i = (\boldsymbol{x}_{1i}, \boldsymbol{z}_{2i})$ are treated as exogenous, where $\boldsymbol{z}_{2i}$ is $\ell_2 \times 1$ and $\ell_2 \geq k_2$. You are interested in testing the hypothesis

$$\mathbb{H}_0 : \boldsymbol{\beta}_2 = 0.$$

Consider the reduced form equation for y_i

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\lambda}_1 + \boldsymbol{z}_{2i}'\boldsymbol{\lambda}_2 + v_i. \tag{11.104}$$

Show how to test $\mathbb{H}_0$ using only the OLS estimates of (11.104).

Hint: This will require an analysis of the reduced form equations and their relation to the structural equation.

Exercise 11.14 Take the linear instrumental variables equation

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= 0 \end{aligned}$$

where $\boldsymbol{x}_{1i}$ is $k_1 \times 1$, $\boldsymbol{x}_{2i}$ is $k_2 \times 1$, and $\boldsymbol{z}_i$ is $\ell \times 1$, with $\ell \geq k = k_1 + k_2$. The sample size is n. Assume that $\boldsymbol{Q}_{\boldsymbol{zz}} = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i'\right) > 0$ and $Q_{\boldsymbol{zx}} = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right)$ has full rank k.

Suppose that only $(y_i, \boldsymbol{x}_{1i}, \boldsymbol{z}_i)$ are available, and $\boldsymbol{x}_{2i}$ is missing from the dataset.

Consider the 2SLS estimator $\widehat{\boldsymbol{\beta}}_1$ of $\boldsymbol{\beta}_1$ obtained from the misspecified IV regression, by regressing y_i on $\boldsymbol{x}_{1i}$ only, using $\boldsymbol{z}_i$ as an instrument for $\boldsymbol{x}_{1i}$.

(a) Find a stochastic decomposition $\widehat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1 + \boldsymbol{b}_{1n} + \boldsymbol{r}_{1n}$ where $\boldsymbol{r}_{1n}$ depends on the error e_i, and $\boldsymbol{b}_{1n}$ does not depend on the error e_i.

(b) Show that $\boldsymbol{r}_{1n} \to_p 0$ as $n \to \infty$.

(c) Find the probability limit of $\boldsymbol{b}_{1n}$ and $\widehat{\boldsymbol{\beta}}_1$ as $n \to \infty$.

(d) Does $\widehat{\boldsymbol{\beta}}_1$ suffer from "omitted variables bias"? Explain. Under what conditions is there no omitted variables bias?

(e) Find the asymptotic distribution as $n \to \infty$ of

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 - \boldsymbol{b}_{1n}\right).$$

Exercise 11.15 Take the linear instrumental variables equation

$$\begin{aligned} y_i &= x_i\beta_1 + z_i\beta_2 + e_i \\ \mathbb{E}\left(e_i|z_i\right) &= 0 \end{aligned}$$

where for simplicity both x_i and z_i are scalar 1×1.

(a) Can the coefficients (β_1, β_2) be estimated by 2SLS using z_i as an instrument for x_i? Why or why not?

(b) Can the coefficients (β_1, β_2) be estimated by 2SLS using z_i and z_i^2 as instruments?

(c) For the 2SLS estimator suggested in (b), what is the implicit exclusion restriction?

(d) In (b), what is the implicit assumption about instrument relevance?
[Hint: Write down the implied reduced form equation for x_i.]

(e) In a generic application, would you be comfortable with the assumptions in (c) and (d)?

Exercise 11.16 Take a linear equation with endogeneity and a just-identified linear reduced form

$$\begin{aligned} y_i &= x_i\beta + e_i \\ x_i &= \gamma z_i + u_i \end{aligned}$$

where both x_i and z_i are scalar 1×1. Assume that

$$\begin{aligned} \mathbb{E}(z_i e_i) &= 0 \\ \mathbb{E}(z_i u_i) &= 0 \end{aligned}$$

(a) Derive the reduced form equation

$$y_i = z_i\lambda + v_i.$$

Show that $\beta = \lambda/\gamma$ if $\gamma \neq 0$, and that $\mathbb{E}(z_i v_i) = 0$

(b) Let $\widehat{\lambda}$ denote the OLS estimate from linear regression of Y on Z, and let $\widehat{\gamma}$ denote the OLS estimate from linear regression of X on Z. Write $\theta = (\lambda, \gamma)'$ and let $\widehat{\theta} = (\widehat{\lambda}, \widehat{\gamma})'$. Define the error vector $\boldsymbol{\xi}_i = \begin{pmatrix} v_i \\ u_i \end{pmatrix}$. Write $\sqrt{n}\left(\widehat{\theta} - \theta\right)$ using a single expression as a function of the error $\boldsymbol{\xi}_i$.

(c) Show that $\mathbb{E}(z_i\boldsymbol{\xi}_i) = 0$

(d) Derive the joint asymptotic distribution of $\sqrt{n}\left(\widehat{\theta}-\theta\right)$ as $n\to\infty$. Hint: Define $\boldsymbol{\Omega}_{\boldsymbol{\xi}}=\mathbb{E}\left(z_i^2\boldsymbol{\xi}_i\boldsymbol{\xi}_i'\right)$

(e) Using the previous result and the Delta Method, find the asymptotic distribution of the Indirect Least Squares estimator $\widehat{\beta}=\widehat{\lambda}/\widehat{\gamma}$

(f) Is the answer in (e) the same as the asymptotic distribution of the 2SLS estimator in Theorem 11.14.1?

Hint: Show that $\begin{pmatrix}1 & -\beta\end{pmatrix}\boldsymbol{\xi}_i=e_i$ and $\begin{pmatrix}1 & -\beta\end{pmatrix}\boldsymbol{\Omega}_{\boldsymbol{\xi}}\begin{pmatrix}1\\-\beta\end{pmatrix}=\mathbb{E}\left(z_i^2e_i^2\right)$.

Exercise 11.17 Take the model

$$\begin{aligned}y_i&=\boldsymbol{x}_i'\boldsymbol{\beta}+e_i\\\mathbb{E}\left(z_ie_i\right)&=0\end{aligned}$$

and consider the two-stage least-squares estimator. The first-stage estimate is

$$\begin{aligned}\widehat{\boldsymbol{X}}&=\boldsymbol{Z}\widehat{\boldsymbol{\Gamma}}\\\widehat{\boldsymbol{\Gamma}}&=\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\end{aligned}$$

and the second-stage is least-squares of y_i on $\widehat{\boldsymbol{x}}_i$:

$$\widehat{\boldsymbol{\beta}}=\left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{y}$$

with least-squares residuals

$$\widehat{\boldsymbol{e}}=\boldsymbol{y}-\widehat{\boldsymbol{X}}\widehat{\boldsymbol{\beta}}.$$

Consider $\widehat{\sigma}^2=\dfrac{1}{n}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$ as an estimator for $\sigma^2=\mathbb{E}\left(e_i^2\right)$. Is this appropriate? If not, propose an alternative estimator.

Exercise 11.18 You have two independent iid samples $(y_{1i},\boldsymbol{x}_{1i},\boldsymbol{z}_{1i}:i=1,...,n)$ and $(y_{2i},\boldsymbol{x}_{2i},\boldsymbol{z}_{2i}:i=1,...,n)$. The dependent variables y_{1i} and y_{2i} are real-valued. The regressors $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ and instruments $\boldsymbol{z}_{1i}$ and $\boldsymbol{z}_{2i}$ are k-vectors. The model is standard just-identified linear instrumental variables

$$\begin{aligned}y_{1i}&=\boldsymbol{x}_{1i}'\boldsymbol{\beta}_1+e_{1i}\\\mathbb{E}\left(\boldsymbol{z}_{1i}e_{1i}\right)&=\mathbf{0}\\y_{2i}&=\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2+e_{2i}\\\mathbb{E}\left(\boldsymbol{z}_{2i}e_{2i}\right)&=\mathbf{0}\end{aligned}$$

For concreteness, sample 1 are women and sample 2 are men. You want to test $\mathbb{H}_0:\boldsymbol{\beta}_1=\boldsymbol{\beta}_2$, that the two samples have the same coefficients.

(a) Develop a test statistic for $\mathbb{H}_0$.

(b) Derive the asymptotic distribution of the test statistic.

(c) Describe (in brief) the testing procedure.

Exercise 11.19 To estimate β in the model $y_i=x_i\beta+e_i$ with x_i scalar and endogenous, with household level data, you want to use as an the instrument the state of residence.

(a) What are the assumptions needed to justify this choice of instrument?

(b) Is the model just identified or overidentified?

Exercise 11.20 The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= \mathbf{0} \end{aligned}$$

An economist wants to obtain the 2SLS estimates and standard errors for $\boldsymbol{\beta}$. He uses the following steps

- Regresses $\boldsymbol{x}_i$ on $\boldsymbol{z}_i$, obtains the predicted values $\widehat{\boldsymbol{x}}_i$.
- Regresses y_i on $\widehat{\boldsymbol{x}}_i$, obtains the coefficient estimate $\widehat{\boldsymbol{\beta}}$ and standard error $s(\widehat{\boldsymbol{\beta}})$ from this regression.

Is this correct? Does this produce the 2SLS estimates and standard errors?

Exercise 11.21 Let

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i$$

Let $(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$ denote the 2SLS estimates of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ when $\boldsymbol{z}_{2i}$ is used as an instrument for $\boldsymbol{x}_{2i}$ and they are the same dimension (so the model is just identified). Let $(\widehat{\boldsymbol{\lambda}}_1, \widehat{\boldsymbol{\lambda}}_2)$ be the OLS estimates from the regression

$$y_i = \boldsymbol{x}_{1i}'\widehat{\boldsymbol{\lambda}}_1 + \boldsymbol{z}_{2i}'\widehat{\boldsymbol{\lambda}}_2 + e_i$$

Show that $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\lambda}}_1$.

Exercise 11.22 In the linear model

$$y_i = x_i\beta + e_i$$

suppose $\sigma_i^2 = E\left(e_i^2 \mid x_i\right)$ is known. Show that the GLS estimator of β can be written as an instrumental variables estimator using some instrument z_i. (Find an expression for z_i.)

Exercise 11.23 You will replicate and extend the work reported in Acemoglu, Johnson and Robinson (2001). The authors provided an expanded set of controls when they published their 2012 extension and posted the data on the AER website. This dataset is `AJR2001` on the textbook website..

(a) Estimate the OLS regression (11.95), the reduced form regression (11.96) and the 2SLS regression (11.97). (Which point estimate is different by 0.01 from the reported values? This is a common phenomenon in empirical replication).

(b) For the above estimates, calculate both homoskedastic and heteroskedastic-robust standard errors. Which were used by the authors (as reported in (11.95)-(11.96)-(11.97)?)

(c) Calculate the 2SLS estimates by the Indirect Least Squares formula. Are they the same?

(d) Calculate the 2SLS estimates by the two-stage approach. Are they the same?

(e) Calculate the 2SLS estimates by the control variable approach. Are they the same?

(f) Acemoglu, Johnson and Robinson (2001) reported many specifications including alternative regressor controls, for example *latitude* and *africa*. Estimate by least-squares the equation for logGDP adding *latitude* and *africa* as regressors. Does this regression suggest that *latitude* and *africa* are predictive of the level of GDP?

(g) Now estimate the same equation as in (f) but by 2SLS using log mortality as an instrument for *risk*. How does the interpretation of the effect of *latitude* and *africa* change?

(h) Return to our baseline model (without including *latitude* and *africa*). The authors' reduced form equation uses log(mortality) as the instrument, rather than, say, the level of mortality. Estimate the reduced form for risk with *mortality* as the instrument. (This variable is not provided in the dataset, so you need to take the exponential of the mortality variable.) Can you explain why the authors preferred the equation with log(mortality)?

(i) Try an alternative reduced form, including both log(mortality) and the square of log(mortality). Interpret the results. Re-estimate the structural equation by 2SLS using both log(mortality) and its square as instruments. How do the results change?

(j) For the estimates in (i), are the instruments strong or weak using the Stock-Yogo test?

(k) Calculate and interpret a test for exogeneity of the instruments.

(l) Estimate the equation by LIML, using the instruments log(mortality) and the square of log(mortality).

Exercise 11.24 You will replicate and extend the work reported in the chapter relating to Card (1995). The data is from the author's website, and is posted as `Card1995`. The model we focus on is labeled 2SLS(a) in Table 11.1, which uses *public* and *private* as instruments for *Edu*. The variables you will need for this exercise include *lwage76*, *ed76* , *age76*, *smsa76r*, *reg76r*, *nearc2*, *nearc4*, *nearc4a*, *nearc4b*. See the description file for definitions.

$$\log(Wage) = \beta_0 + \beta_1 Edu + \beta_2 Exp + \beta_3 Exp^2/100 + \beta_4 South + \beta_5 Black + \boldsymbol{e}$$

where $Edu = Education$ (Years), $Exp = Experience$ (Years), and $South$ and $Black$ are regional and racial dummy variables. The variables $Exp = Age - Edu - 6$ and $Exp^2/100$ are not in the dataset, they need to be generated.

(a) First, replicate the reduced form regression presented in the final column of Table 11.2, and the 2SLS regression described above (using *public* and *private* as instruments for *Edu*) to verify that you have the same variable defintions.

(b) Now try a different reduced form model. The variable *nearc2* means "grew up near a 2-year college". See if adding it to the reduced form equation is useful.

(c) Now try more interactions in the reduced form. Create the interactions *nearc4a*age76* and *nearc4a*age76*$^2/100$, and add them to the reduced form equation. Estimate this by least-squares. Interpret the coefficients on the two new variables.

(d) Estimate the structural equation by 2SLS using the expanded instrument set

{*nearc4a, nearc4b, nearc4a*age76, nearc4a*age76*$^2/100$}.

What is the impact on the structural estimate of the return to schooling?

(e) Using the Stock-Yogo test, are the instruments strong or weak?

(f) Test the hypothesis that *Edu* is exogenous for the structural return to schooling.

(g) Re-estimate the last equation by LIML. Do the results change meaningfully?

Exercise 11.25 You will extend Angrist and Krueger (1991). In their Table VIII, they report their estimates of an analog of (11.99) for the subsample of 26,913 black men. Use this sub-sample for the following analysis.

(a) Start by considering estimation of an equation which is identical in form to (11.99), with the same additional regressors (year-of-birth, region-of-residence, and state-of-birth dummy variables) and 180 excluded instrumental variables (the interactions of quarter-of-birth times year-of-birth dummy variables, and quarter-of-birth times state-of-birth interactions). But now, it is estimated on the subsample of black men. One regressor must be omitted to achieve identification. Which variable is this?

(b) Estimate the reduced form for the above equation by least-squares. Calculate the F statistic for the excluded instruments. What do you conclude about the strength of the instruments?

(c) Repeat, now estimating the reduced form for the analog of (11.98) which has 30 excluded instrumental variables, and does not include the state-of-birth dummy variables in the regression. What do you conclude about the strength of the instruments?

(d) Repeat, now estimating the reduced form for the analog of (11.101) which has only 3 excluded instrumental variables. Are the instruments sufficiently strong for 2SLS estimation? For LIML estimation?

(e) Estimate the structural wage equation using what you believe is the most appropriate set of regressors, instruments, and the most appropriate estimation method. What is the estimated return to education (for the subsample of black men) and its standard error? Without doing a formal hypothesis test, do these results (or in which way?) appear meaningfully different from the results for the full sample?

Chapter 12

Generalized Method of Moments

12.1 Moment Equation Models

All of the models that have been introduced so far can be written as **moment equation models**, where the population parameters solve a system of moment equations. Moment equation models are much broader than the models so far considered, and understanding their common structure opens up straightforward techniques to handle new econometric models.

Moment equation models take the following form. Let $\boldsymbol{g}_i(\boldsymbol{\beta})$ be a known $\ell \times 1$ function of the i^{th} observation and a $k \times 1$ parameter $\boldsymbol{\beta}$. A moment equation model is summarized by the moment equations

$$\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\right) = \mathbf{0} \tag{12.1}$$

and a parameter space $\boldsymbol{\beta} \in \boldsymbol{B}$. For example, in the instrumental variables model $\boldsymbol{g}_i\left(\boldsymbol{\beta}\right) = \boldsymbol{z}_i\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)$.

In general, we say that a parameter $\boldsymbol{\beta}$ is **identified** if there is a unique mapping from the data distribution to $\boldsymbol{\beta}$. In the context of the model (12.1) this means that there is a unique $\boldsymbol{\beta}$ satisfying (12.1). Since (12.1) is a system of ℓ equations with k unknowns, then it is necessary that $\ell \geq k$ for there to be a unique solution. If $\ell = k$ we say that the model is **just identified**, meaning that there is just enough information to identify the parameters. If $\ell > k$ we say that the model is **overidentified**, meaning that there is excess information (which can improve estimation efficiency). If $\ell < k$ we say that the model is **underidentified**, meaning that there is insufficient information to identify the parameters. In general, we assume that $\ell \geq k$ so the model is either just identified or overidentified.

12.2 Method of Moments Estimators

In this section we consider the just-identified case $\ell = k$.

Define the sample analog of (12.5)

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\beta}). \tag{12.2}$$

The **method of moments estimator (MME)** $\widehat{\boldsymbol{\beta}}_{\text{mm}}$ for $\boldsymbol{\beta}$ is defined as the parameter value which sets $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \mathbf{0}$. Thus

$$\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}_{\text{mm}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\widehat{\boldsymbol{\beta}}_{\text{mm}}) = \mathbf{0}. \tag{12.3}$$

The equations (12.3) are known as the **estimating equations** as they are the equations which determine the estimator $\widehat{\boldsymbol{\beta}}_{\text{mm}}$.

In some contexts (such as those discussed in the examples below), their is an explicit solution for $\widehat{\boldsymbol{\beta}}_{\text{mm}}$. In other cases the solution must be found numerically.

We now show how most of the estimators discussed so far in the textbook can be written as method of moments estimators.

Mean: Set $g_i(\mu) = y_i - \mu$. The MME is $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n y_i$.

Mean and Variance: Set

$$\boldsymbol{g}_i\left(\mu, \sigma^2\right) = \begin{pmatrix} y_i - \mu \\ (y_i - \mu)^2 - \sigma^2 \end{pmatrix}.$$

The MME are $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \widehat{\mu})^2$.

OLS: Set $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$. The MME is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y})$.

OLS and Variance: Set

$$\boldsymbol{g}_i\left(\boldsymbol{\beta}, \sigma^2\right) = \begin{pmatrix} \boldsymbol{x}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) \\ (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 - \sigma^2 \end{pmatrix}.$$

The MME is $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{y})$ and $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right)^2$.

Multivariate Least Squares, vector form: Set $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{X}_i(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta})$. The MME is $\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i'\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{y}_i\right)$ which is (10.3).

Multivariate Least Squares, matrix form: Set $\boldsymbol{g}_i(\boldsymbol{B}) = \operatorname{vec}(\boldsymbol{x}_i(\boldsymbol{y}_i' - \boldsymbol{x}_i'\boldsymbol{B}))$. The MME is $\widehat{\boldsymbol{B}} = \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{y}_i'\right)$ which is (10.5).

Seemingly Unrelated Regression: Set

$$\boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{\Sigma}) = \begin{pmatrix} \boldsymbol{X}_i\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta}) \\ \operatorname{vec}\left(\boldsymbol{\Sigma} - (\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta})(\boldsymbol{y}_i - \boldsymbol{X}_i'\boldsymbol{\beta})'\right) \end{pmatrix}.$$

The MME is $\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \boldsymbol{X}_i\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}_i'\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{X}_i\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{y}_i\right)$ and $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \left(\boldsymbol{y}_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}\right)\left(\boldsymbol{y}_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}\right)'$.

IV: Set $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta})$. The MME is $\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{z}_i y_i\right)$.

Generated Regressors: Set

$$\boldsymbol{g}_i(\boldsymbol{\beta}, \boldsymbol{A}) = \begin{pmatrix} \boldsymbol{A}'\boldsymbol{z}_i(y_i - \boldsymbol{z}_i'\boldsymbol{A}\boldsymbol{\beta}) \\ \operatorname{vec}(\boldsymbol{z}_i(\boldsymbol{x}_i' - \boldsymbol{z}_i'\boldsymbol{A})) \end{pmatrix}.$$

The MME is $\widehat{\boldsymbol{A}} = \left(\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i'\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{x}_i'\right)$ and $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{A}}'\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i'\widehat{\boldsymbol{A}}\right)^{-1}\left(\widehat{\boldsymbol{A}}'\sum_{i=1}^n \boldsymbol{z}_i y_i\right)$.

A common feature unifying these examples is that the estimator can be written as the solution to a set of estimating equations (12.3). This provides a common framework which enables a convenient development of a unified distribution theory.

12.3 Overidentified Moment Equations

In the instrumental variables model $g_i(\boldsymbol{\beta}) = \boldsymbol{z}_i\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)$. Thus (12.2) is

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right) = \frac{1}{n}\left(\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta}\right). \tag{12.4}$$

We have defined the method of moments estimator for $\boldsymbol{\beta}$ as the parameter value which sets $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \mathbf{0}$. However, when the model is overidentified (if $\ell > k$) then this is generally impossible as there are more equations than free parameters. Equivalently, there is no choice of $\boldsymbol{\beta}$ which sets (12.4) to zero. Thus the method of moments estimator is not defined for the overidentified case.

While we cannot find an estimator which sets $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$ equal to zero, can can try to find an estimator which makes $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$ as close to zero as possible. Let's think what that means. Since $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$ is an $\ell \times 1$ vector, this means we are trying to find a value for $\boldsymbol{\beta}$ which sets $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$ as close as possible to the zero vector.

One way to think about this is to define the vector $\boldsymbol{\mu} = \boldsymbol{Z}'\boldsymbol{y}$, the matrix $\boldsymbol{G} = \boldsymbol{Z}'\boldsymbol{X}$ and the "error" $\boldsymbol{\eta} = \boldsymbol{\mu} - \boldsymbol{G}\boldsymbol{\beta}$. Then we can write (12.4) as

$$\boldsymbol{\mu} = \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{\eta}.$$

This looks like a regression equation with the $\ell \times 1$ dependent variable $\boldsymbol{\mu}$, the $\ell \times k$ regressor matrix $\boldsymbol{G}$, and the $\ell \times 1$ error vector $\boldsymbol{\eta}$. Recall, the goal is to make the error vector $\boldsymbol{\eta}$ as small as possible. Recalling our knowledge about least-squares, we know that a simple method is to use least-squares regression of $\boldsymbol{\mu}$ on $\boldsymbol{G}$, which minimzes the sum-of-squares $\boldsymbol{\eta}'\boldsymbol{\eta}$. This is certainly one way to make $\boldsymbol{\eta}$ "small". This least-squares solution is $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{G}'\boldsymbol{G}\right)^{-1}\left(\boldsymbol{G}'\boldsymbol{\mu}\right)$.

More generally, we know that when errors are non-homogeneous it can be more efficient to estimate by weighted least squares. Thus for some weight matrix $\boldsymbol{W}$, consider the estimator

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{G}'\boldsymbol{W}\boldsymbol{G}\right)^{-1}\left(\boldsymbol{G}'\boldsymbol{W}\boldsymbol{\mu}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{y}\right).\end{aligned}$$

This minimizes the weighted sum of squares $\boldsymbol{\eta}'\boldsymbol{W}\boldsymbol{\eta}$. This solution is known as the generalized method of moments (GMM).

The estimator is typically defined as follows. Given a set of moment equations (12.2) and an $\ell \times \ell$ weight matrix $\boldsymbol{W} > 0$, the GMM criterion function is defined as

$$J(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})'\boldsymbol{W}\,\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}).$$

The factor "n" is not important for the definition of the estimator, but is convenient for the distribution theory. The criterion $J(\boldsymbol{\beta})$ is the weighted sum of squared moment equation errors. When $\boldsymbol{W} = \boldsymbol{I}_\ell$, then $J(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})'\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = n \cdot \left\|\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})\right\|^2$, the square of the Euclidean length. Since we restrict attention to positive definite weight matrices $\boldsymbol{W}$, the criterion $J(\boldsymbol{\beta})$ is always non-negative.

The **Generalized Method of Moments (GMM)** estimator is defined as the minimizer of the GMM criterion $J(\boldsymbol{\beta})$.

Definition 12.3.1 *The Generalized Method of Moments estimator is* $\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, J_n\left(\boldsymbol{\beta}\right)$.

Recall that in the just-identified case $k = \ell$, the method of moments estimator $\widehat{\boldsymbol{\beta}}_{\text{mm}}$ solves $\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}_{\text{mm}}) = \mathbf{0}$. Hence in this case $J_n\left(\widehat{\boldsymbol{\beta}}_{\text{mm}}\right) = 0$ which means that $\widehat{\boldsymbol{\beta}}_{\text{mm}}$ minimizes $J_n(\boldsymbol{\beta})$ and equals $\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \widehat{\boldsymbol{\beta}}_{\text{mm}}$. This means that GMM includes MME as a special case. This implies that all of our results for GMM will apply to any method of moments estimators as a special case.

In the over-identified case the GMM estimator will depend on the choice of weight matrix $\boldsymbol{W}$ and so this is an important focus of the theory. In the just-identified case, the GMM estimator simplifies to the MME which does not depend on $\boldsymbol{W}$.

The method and theory of the generalized method of moments was developed in an influential paper by Lars Hansen (1982). This paper introduced the method, its asymptotic distribution, the form of the efficient weight matrix, and tests for overidentification.

Lars Peter Hansen

Lars Hansen (1952-) is an American econometrician and macroeconomist. In econometrics, he is famously known for the GMM estimator which has transformed theoretical and empirical economics. He was awarded the Nobel Memorial Prize in Economics in 2013.

12.4 Linear Moment Models

One of the great advantages of the moment equation framework is that it allows both linear and nonlinear models. However, when the moment equations are linear in the parameters then we have explicit solutions for the estimates and a straightforward asymptotic distribution theory. Hence we start by confining attention to linear moment equations, and return to nonlinear moment equations later. In the examples listed earlier, the estimators which have linear moment equations include the sample mean, OLS, multivariate least squares, IV, and 2SLS. The estimates which have non-linear moment equations include the sample variance, SUR, and generated regressors.

In particular, we focus on the overidentified IV model

$$\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) \tag{12.5}$$

where $\boldsymbol{z}_i$ is $\ell \times 1$ and $\boldsymbol{x}_i$ is $k \times 1$.

12.5 GMM Estimator

Given (12.5) the sample moment equations are (12.4). The GMM criterion can be written as

$$J(\boldsymbol{\beta}) = n\left(\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta}\right)' \boldsymbol{W}\left(\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta}\right).$$

The GMM estimator minimizes $J(\boldsymbol{\beta})$. The first order conditions are

$$\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\beta}} J(\widehat{\boldsymbol{\beta}}) \\
&= 2\frac{\partial}{\partial \boldsymbol{\beta}} \overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})' \boldsymbol{W} \overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}) \\
&= -2\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right) \boldsymbol{W}\left(\frac{1}{n}\boldsymbol{Z}'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\right).
\end{aligned}$$

The solution is given as follows.

Theorem 12.5.1 *For the overidentified IV model*

$$\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{y}\right). \tag{12.6}$$

While the estimator depends on $\boldsymbol{W}$, the dependence is only up to scale. This is because if $\boldsymbol{W}$ is replaced by $c\boldsymbol{W}$ for some $c > 0$, $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ does not change.

When $\boldsymbol{W}$ is fixed by the user, we call $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ a **one-step GMM** estimator.

The GMM estimator (12.6) resembles the 2SLS estimator (11.34). In fact they are equal when $\boldsymbol{W} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. This means that the 2SLS estimator is a one-step GMM estimator for the linear model. In the just-identified case it also simplifies to the IV estimator (11.29).

Theorem 12.5.2 *If* $\boldsymbol{W} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ *then* $\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \widehat{\boldsymbol{\beta}}_{\text{2sls}}$.
Furthermore, if $k = \ell$ *then* $\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \widehat{\boldsymbol{\beta}}_{\text{iv}}$.

12.6 Distribution of GMM Estimator

Let

$$\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i'\right)$$

and

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right) = \mathbb{E}\left(\boldsymbol{g}_i\boldsymbol{g}_i'\right)$$

where $\boldsymbol{g}_i = \boldsymbol{z}_i e_i$. Then

$$\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{W}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right) \xrightarrow{p} \boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}$$

and

$$\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{W}\left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\boldsymbol{e}\right) \xrightarrow{d} \boldsymbol{Q}'\boldsymbol{W}\cdot \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right).$$

We conclude:

Theorem 12.6.1 ***Asymptotic Distribution of GMM Estimator.***
Under Assumption 11.14.1, as $n \to \infty$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Q}\right)\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}. \tag{12.7}$$

We find that the GMM estimator is asymptotically normal with a "sandwich form" asymptotic variance.

Our derivation treated the weight matrix $\boldsymbol{W}$ as if it is non-random, but Theorem 12.6.1 carries over to the case where the weight matrix $\widehat{\boldsymbol{W}}$ is random so long as it converges in probability to some positive definite limit $\boldsymbol{W}$. This may require scaling the weight matrix, for example replacing $\widehat{\boldsymbol{W}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ with $\widehat{\boldsymbol{W}} = \left(n^{-1}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$. Since rescaling the weight matrix does not affect the estimator this is ignored in implementation.

12.7 Efficient GMM

The asymptotic distribution of the GMM estimator $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ depends on the weight matrix $\boldsymbol{W}$ through the asymptotic variance $\boldsymbol{V}_{\boldsymbol{\beta}}$. The asymptotically optimal weight matrix $\boldsymbol{W}_0$ is one which minimizes $\boldsymbol{V}_{\boldsymbol{\beta}}$. This turns out to be $\boldsymbol{W}_0 = \boldsymbol{\Omega}^{-1}$. The proof is left to Exercise 12.4.

When the GMM estimator $\widehat{\boldsymbol{\beta}}$ is constructed with $\boldsymbol{W} = \boldsymbol{W}_0 = \boldsymbol{\Omega}^{-1}$ (or a weight matrix which is a consistent estimator of $\boldsymbol{W}_0$) we call it the **Efficient GMM** estimator:

$$\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \left(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Omega}^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Omega}^{-1}\boldsymbol{Z}'\boldsymbol{y}\right).$$

Its asymptotic distribution takes a simpler form than in Theorem 12.6.1. By substituting $\boldsymbol{W} = \boldsymbol{W}_0 = \boldsymbol{\Omega}^{-1}$ into (12.7) we find

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}\left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}\boldsymbol{\Omega}^{-\mathbf{1}}\boldsymbol{Q}\right)\left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1} = \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}.$$

This is the asymptotic variance of the efficient GMM estimator.

Theorem 12.7.1 ***Asymptotic Distribution of GMM with Efficient Weight Matrix.*** *Under Assumption 11.14.1 and $\boldsymbol{\Omega} > 0$, as $n \to \infty$*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}.$$

Theorem 12.7.2 ***Efficient GMM.*** *Under Assumption 11.14.1 and $\boldsymbol{\Omega} > 0$, for any $\boldsymbol{W} > 0$,*

$$\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Q}\right)\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1} - \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1} > 0.$$

Thus if $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ is the efficient GMM estimator and $\widetilde{\boldsymbol{\beta}}_{\text{gmm}}$ is another GMM estimator, then

$$\operatorname{avar}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}}\right) \leq \operatorname{avar}\left(\widetilde{\boldsymbol{\beta}}_{\text{gmm}}\right).$$

For a proof, see Exercise 12.4.

This means that the smallest possible GMM covariance matrix (in the positive definite sense) is achieved by the efficient GMM weight matrix.

$\boldsymbol{W}_0 = \boldsymbol{\Omega}^{-1}$ is not known in practice but it can be estimated consistently as we discuss in Section 12.9. For any $\widehat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{W}_0$, the asymptotic distribution in Theorem 12.7.1 is unaffected. Consequently we still call any $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ constructed with an estimate of the efficient weight matrix an efficient GMM estimator.

By "efficient", we mean that this estimator has the smallest asymptotic variance in the class of GMM estimators with this set of moment conditions. This is a weak concept of optimality, as we are only considering alternative weight matrices $\widehat{\boldsymbol{W}}$. However, it turns out that the GMM estimator is semiparametrically efficient as shown by Gary Chamberlain (1987). If it is known that $\mathbb{E}\left(\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta})\right) = \mathbf{0}$, and this is all that is known, this is a semi-parametric problem as the

distribution of the data is unknown. Chamberlain showed that in this context no semiparametric estimator (one which is consistent globally for the class of models considered) can have a smaller asymptotic variance than $\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}$ where $\boldsymbol{G} = \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\boldsymbol{\beta})\right)$. Since the GMM estimator has this asymptotic variance, it is semiparametrically efficient.

The results in this section show that in the linear model no estimator has better asymptotic efficiency than the efficient linear GMM estimator. No estimator can do better (in this first-order asymptotic sense), without imposing additional assumptions.

12.8 Efficient GMM versus 2SLS

For the linear model we introduced the 2SLS estimator as a standard estimator for $\boldsymbol{\beta}$. Now we have introduced the GMM estimator which includes 2SLS as a special case. Is there a context where 2SLS is efficient?

To answer this question, recall that the 2SLS estimator is GMM given the weight matrix $\widehat{\boldsymbol{W}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ or equivalently $\widehat{\boldsymbol{W}} = \left(n^{-1}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$ since scaling doesn't matter. Since $\widehat{\boldsymbol{W}} \xrightarrow{p} \left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}$, this is asymptotically equivalent to using the weight matrix $\boldsymbol{W} = \left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}$. In contrast, the efficient weight matrix takes the form $\left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)\right)^{-1}$. Now suppose that the structural equation error e_i is conditionally homoskedastic in the sense that $\mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right) = \sigma^2$. Then the efficient weight matrix equals $\boldsymbol{W} = \left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}\sigma^{-2}$, or equivalently $\boldsymbol{W} = \left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}$ since scaling doesn't matter. The latter weight matrix is the same as the 2SLS asymptotic weight matrix. This shows that the 2SLS weight matrix is the efficient weight matrix under conditional homoskedasticity.

Theorem 12.8.1 *Under Assumption 11.14.1 and* $\mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right) = \sigma^2$ *then* $\widehat{\boldsymbol{\beta}}_{2\text{sls}}$ *is efficient GMM.*

This shows that 2SLS is efficient under homoskedasticity. When homoskedasticity holds, there is no reason to use efficient GMM over 2SLS. More broadly, when homoskedasticity is a reasonable approximation then 2SLS will be a reasonable estimator. However, this result also shows that in the general case where the error is conditionally heteroskedastic, then 2SLS is generically inefficient relative to efficient GMM.

12.9 Estimation of the Efficient Weight Matrix

To construct the efficient GMM estimator we need a consistent estimator $\widehat{\boldsymbol{W}}$ of $\boldsymbol{W}_0 = \boldsymbol{\Omega}^{-1}$. The convention is to form an estimate $\widehat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ and then set $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{\Omega}}^{-1}$.

The **two-step GMM estimator** proceeds by using a one-step consistent estimate of $\boldsymbol{\beta}$ to construct the weight matrix estimator $\widehat{\boldsymbol{W}}$. In the linear model the natural one-step estimator for $\boldsymbol{\beta}$ is the 2SLS estimator $\widehat{\boldsymbol{\beta}}_{2\text{sls}}$. Set $\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{2\text{sls}}$, $\widetilde{\boldsymbol{g}}_i = \boldsymbol{g}_i(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{z}_i\widetilde{e}_i$ and $\overline{\boldsymbol{g}}_n = n^{-1}\sum_{i=1}^n \widetilde{\boldsymbol{g}}_i$. Two moment estimators of $\boldsymbol{\Omega}$ are then

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^n \widetilde{\boldsymbol{g}}_i\widetilde{\boldsymbol{g}}_i' \tag{12.8}$$

and

$$\widehat{\boldsymbol{\Omega}}^* = \frac{1}{n}\sum_{i=1}^n \left(\widetilde{\boldsymbol{g}}_i - \overline{\boldsymbol{g}}_n\right)\left(\widetilde{\boldsymbol{g}}_i - \overline{\boldsymbol{g}}_n\right)'. \tag{12.9}$$

The estimator (12.8) is an uncentered covariance matrix estimator while the estimator (12.9) is a centered version. Either estimator is consistent when $\mathbb{E}\left(\boldsymbol{z}_i e_i\right) = \boldsymbol{0}$ which holds under correct

specification. However under misspecification we may have $\mathbb{E}(\boldsymbol{z}_i e_i) \neq \mathbf{0}$. In the latter context $\widehat{\boldsymbol{\Omega}}^*$ may be viewed as a robust estimator. For some testing problems it turns out to be preferable to use a covariance matrix estimator which is robust to the alternative hypothesis. For these reasons estimator (12.9) is generally preferred. Unfortunately, estimator (12.8) is more commonly seen in practice since it is the default choice by most packages. It is also worth observing that when the model is just identified then $\overline{\boldsymbol{g}}_n = \mathbf{0}$ so the two are algebraically identically.

Given the choice of covariance matrix estimator we set $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{\Omega}}^{-1}$ or $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{\Omega}}^{*-1}$. Given this weight matrix, we then construct the **two-step GMM estimator** as (12.6) using the weight matrix $\widehat{\boldsymbol{W}}$.

Since the 2SLS estimator is consistent for $\boldsymbol{\beta}$, by arguments nearly identical to those used for covariance matrix estimation, we can show that $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\boldsymbol{\Omega}}^*$ are consistent for $\boldsymbol{\Omega}$ and thus $\widehat{\boldsymbol{W}}$ is consistent for $\boldsymbol{\Omega}^{-1}$. See Exercise 12.3.

This also means that the two-step GMM estimator satisfies the conditions for Theorem 12.7.1. We have established.

Theorem 12.9.1 *Under Assumption 11.14.1 and* $\boldsymbol{\Omega} > 0$*, if* $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{\Omega}}^{-1}$ *or* $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{\Omega}}^{*-1}$ *where the latter are defined in (12.8) and (12.9) then as* $n \to \infty$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}.$$

This shows that the two-step GMM estimator is asymptotically efficient.

The two-step GMM estimator of the IV regression equation can be computed in Stata using the `ivregress gmm` command. By default it uses formula (12.8). The centered version (12.9) may be selected using the `center` option.

12.10 Iterated GMM

The asymptotic distribution of the two-step GMM estimator does not depend on the choice of the preliminary one-step estimator. However, the actual value of the estimator depends on this choice, and so will the finite sample distribution. This is undesirable and likely inefficient. To remove this dependence we can iterate the estimation sequence. Specifically, given $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ we can construct an updated weight matrix estimate $\widehat{\boldsymbol{W}}$ and then re-estimate $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$. This updating can be iterated until convergence[1]. The result is called the **iterated GMM estimator** and is a common implementation of efficient GMM.

Interestingly, B. Hansen and Lee (2018) show that the iterated GMM estimator is unaffected if the weight matrix is computed with or without centering. Standard errors and test statistics, however, will be affected by the choice.

The iterated GMM estimator of the IV regression equation can be computed in Stata using the `ivregress gmm` command using the `igmm` option.

[1] In practice, "convergence" obtains when the difference between the estimates obtained at subsequent steps is smaller than a pre-specified tolerance. A sufficient condition for convergence is that the sequence is a contraction mapping. Indeed, B. Hansen and Lee (2018) have shown that the iterated GMM estimator generally satisfies this condition in large samples.

12.11 Covariance Matrix Estimation

An estimator of the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ can be obtained by replacing the matrices in the asymptotic variance formula by consistent estimates.

For the one-step GMM estimator the covariance matrix estimator is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{Q}}\right)^{-1}\left(\widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{W}}\widehat{\boldsymbol{Q}}\right)\left(\widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{Q}}\right)^{-1}$$

where

$$\widehat{\boldsymbol{Q}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{x}_i'$$

and using either the uncentered estimator (12.8) or centered estimator (12.9) with the residuals $\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{gmm}}$.

For the two-step or iterated gmm estimator the covariance matrix estimator is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{Q}}\right)^{-1} = \left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\widehat{\boldsymbol{\Omega}}^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}. \tag{12.10}$$

Again, $\widehat{\boldsymbol{\Omega}}$ can be computed using either the uncentered estimator (12.8) or centered estimator (12.9), but should use the final residuals $\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{gmm}}$.

Asymptotic standard errors are given by the square roots of the diagonal elements of $n^{-1}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$.

In Stata, the default covariance matrix estimation method is determined by the choice of weight matrix. Thus if the centered estimator (12.9) is used for the weight matrix, it is also used for the covariance matrix estimator.

12.12 Clustered Dependence

In Section 4.20 we introduced clustered dependence and in Section 11.21 described covariance matrix estimation for 2SLS. The methods extend naturally to GMM, but with the additional complication of potentially altering weight matrix calculation.

As before, the structural equation for the g^{th} cluster can be written as the matrix system

$$\boldsymbol{y}_g = \boldsymbol{X}_g\boldsymbol{\beta} + \boldsymbol{e}_g.$$

Using this notation the centered GMM estimator with weight matrix $\boldsymbol{W}$ can be written as

$$\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\left(\sum_{g=1}^{G}\boldsymbol{Z}_g'\boldsymbol{e}_g\right).$$

The cluster-robust covariance matrix estimator for $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ is then

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\widehat{\boldsymbol{S}}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1} \tag{12.11}$$

with

$$\widehat{\boldsymbol{S}} = \sum_{g=1}^{G}\boldsymbol{Z}_g'\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{Z}_g \tag{12.12}$$

and the clustered residuals

$$\widehat{\boldsymbol{e}}_g = \boldsymbol{y}_g - \boldsymbol{X}_g\widehat{\boldsymbol{\beta}}_{\text{gmm}}. \tag{12.13}$$

The cluster-robust estimator (12.11) is appropriate for the one-step GMM estimator. It is also appropriate for the two-step and iterated estimators when the latter use a conventional (non-clustered) efficient weight matrix. However in the clustering context it is more natural to use a

cluster-robust weight matrix such as $\boldsymbol{W} = \widehat{\boldsymbol{S}}^{-1}$ where $\widehat{\boldsymbol{S}}$ is a cluster-robust covariance estimator as in (12.12) based on a one-step or iterated residual. This gives rise to the cluster-robust GMM estimator

$$\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \left(\boldsymbol{X}'\boldsymbol{Z}\widehat{\boldsymbol{S}}^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\widehat{\boldsymbol{S}}^{-1}\boldsymbol{Z}'\boldsymbol{y}. \tag{12.14}$$

For this estimator an appropriate cluster-robust covariance matrix estimator is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{Z}\widehat{\boldsymbol{S}}^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}$$

where $\widehat{\boldsymbol{S}}$ is calculated using the final residuals.

To implement a cluster-robust weight matrix, use the 2SLS estimator for first step estimator. Compute the cluster residuals (12.13) and covariance matrix (12.12). Then (12.14) is the two-step GMM estimator. Updating the residuals and covariance matrix, we can iterate the sequence to obtain the iterated GMM estimator.

In Stata, using the `ivregress gmm` command with the `cluster` option implements the two-step GMM estimator using the cluster-robust weight matrix and cluster-robust covariance matrix estimator. To use the centered covariance matrix use the `center` option, and to implement the iterated GMM estimator use the `igmm` option. Alternatively, you can use the `wmatrix` and `vce` options to separately specify the weight matrix and covariance matrix estimation methods.

12.13 Wald Test

For a given function $\boldsymbol{r}(\boldsymbol{\beta}) : \mathbb{R}^k \to \Theta \subset \mathbb{R}^q$ we define the parameter $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$. The GMM estimator of $\boldsymbol{\theta}$ is $\widehat{\boldsymbol{\theta}}_{\text{gmm}} = \boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}}\right)$. By the delta method it is asymptotically normal with covariance matrix

$$\begin{aligned}\boldsymbol{V}_{\boldsymbol{\theta}} &= \boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R} \\ \boldsymbol{R} &= \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}(\boldsymbol{\beta})'.\end{aligned}$$

An estimator of the asymptotic covariance matrix is

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} &= \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}} \\ \widehat{\boldsymbol{R}} &= \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}(\widehat{\boldsymbol{\beta}}_{\text{gmm}})'.\end{aligned}$$

When θ is scalar then an asymptotic standard error for $\widehat{\theta}_{\text{gmm}}$ is formed as $\sqrt{n^{-1}\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}}$.

A standard test of the hypothesis

$$\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

against

$$\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$$

is based on the Wald statistic

$$W = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).$$

Let $G_q(u)$ denote the χ^2_q distribution function.

Theorem 12.13.1 *Under Assumption 11.14.1 and $\boldsymbol{\Omega} > 0$, if $\boldsymbol{r}(\boldsymbol{\beta})$ is continuously differentiable at $\boldsymbol{\beta}$, and $\mathbb{H}_0$ holds, then as $n \to \infty$,*

$$W \xrightarrow{d} \chi^2_q.$$

For c satisfying $\alpha = 1 - G_q(c)$,

$$\Pr(W > c \mid \mathbb{H}_0) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ if $W > c$" has asymptotic size α.

In Stata, the commands `test` and `testparm` can be used after `ivregress gmm` to implement Wald tests of linear hypotheses. The commands `nlcom` and `testnl` can be used after `ivregress gmm` to implement Wald tests of nonlinear hypotheses.

12.14 Restricted GMM

It is often desirable to impose restrictions on the coefficients. In this section we consider estimation subject to the constraints $\boldsymbol{r}(\boldsymbol{\beta}) = \mathbf{0}$.

The constrained GMM estimator minimizes the GMM criterion subject to the constraint. It is defined as

$$\widehat{\boldsymbol{\beta}}_{\text{cgmm}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmin}} J(\boldsymbol{\beta}).$$

This is the parameter vector which makes the estimating equations as close to zero as possible with respect to the weighted quadratic distance while imposing the restriction on the parameters.

It is useful to separately consider the cases wheres $\boldsymbol{r}(\boldsymbol{\beta})$ are linear and nonlinear.

First let's consider the linear case, where $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{c}$. Using the methods of Chapter 8 it is straightforward to derive that given any weight matrix $\boldsymbol{W}$ the constrained GMM estimator is

$$\widehat{\boldsymbol{\beta}}_{\text{cgmm}} = \widehat{\boldsymbol{\beta}}_{\text{gmm}} - \left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{c}\right). \tag{12.15}$$

In particular, when the efficient weight matrix $\boldsymbol{W} = \widehat{\boldsymbol{\Omega}}^{-1}$ is used the constrained GMM estimator can be written as

$$\widehat{\boldsymbol{\beta}}_{\text{cgmm}} = \widehat{\boldsymbol{\beta}}_{\text{gmm}} - \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{c}\right) \tag{12.16}$$

which is the same formula (8.28) as efficient minimum distance.

To derive the asymptotic distribution under the assumption that the restriction is true, make the substitution $\boldsymbol{c} = \boldsymbol{R}'\boldsymbol{\beta}$ in (12.15) to find

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{cgmm}} - \boldsymbol{\beta}\right) = \left(\boldsymbol{I}_k - \left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right). \tag{12.17}$$

which is a linear function of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right)$. Since the asymptotic distribution of the latter is known, it is straightforward to derive that of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{cgmm}} - \boldsymbol{\beta}\right)$. We present the result for the efficient case in Theorem 12.14.1 below.

Second, let's consider the nonlinear case, meaning that $\boldsymbol{r}(\boldsymbol{\beta})$ is not an affine function of $\boldsymbol{\beta}$. In this case there is (in general) no explicit solution for $\widehat{\boldsymbol{\beta}}_{\text{cgmm}}$. Instead, the solution needs to be found numerically. Fortunately there are excellent nonlinear constrainted optimization solvers which make the task quite feasible. We do not review these here, but can be found in any numerical software system.

For the asymptotic distribution assume again that the restriction $\boldsymbol{r}(\boldsymbol{\beta}) = \mathbf{0}$ is true. Then, using the same methods as in the proof of Theorem 8.14.1 we can show that (12.15) approximately holds, in the sense that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{cgmm}} - \boldsymbol{\beta}\right) = \left(\boldsymbol{I}_k - \left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{Z}\,\boldsymbol{W}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right) + o_p(1) \tag{12.18}$$

where $\boldsymbol{R} = \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}(\boldsymbol{\beta})'$. Thus the asymptotic distribution of the constrained estimator takes the same form as in the linear case.

Theorem 12.14.1 *Under Assumptions 11.14.1 and 8.14.1, and* $\boldsymbol{\Omega} > 0$*, for the efficient constrained GMM estimator (12.16)*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{cgmm}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\text{cgmm}}\right)$$

as $n \to \infty$*, where*

$$\boldsymbol{V}_{\text{cgmm}} = \boldsymbol{V}_{\boldsymbol{\beta}} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}.$$

The asymptotic covariance matrix is estimated by

$$\begin{aligned}
\widehat{\boldsymbol{V}}_{\text{cgmm}} &= \widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}} - \widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}. \\
\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}} &= \left(\widehat{\boldsymbol{Q}}'\widetilde{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{Q}}\right)^{-1} \\
\widetilde{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i\boldsymbol{z}_i'\widetilde{e}_i^2 \\
\widetilde{e}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{\text{cgmm}} \\
\widehat{\boldsymbol{R}} &= \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{r}\left(\widehat{\boldsymbol{\beta}}_{\text{cgmm}}\right)'.
\end{aligned}$$

12.15 Constrained Regression

Take the conventional projection model

$$\begin{aligned}
y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\
\mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0}.
\end{aligned}$$

We can view this as a very special case of GMM. It is model (12.5) with $\boldsymbol{z}_i = \boldsymbol{x}_i$. This is just-identified GMM and the estimator is least-squares $\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \widehat{\boldsymbol{\beta}}_{\text{ols}}$.

In Chapter 8 we discussed estimation of the projection model subject to linear constraints $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}$, which includes exclusion restrictions. Since the projection model is a special case of GMM, the constrained projection model is also constrained GMM. From the results of the previous section we find that the efficient constrained GMM estimator is

$$\widehat{\boldsymbol{\beta}}_{\text{cgmm}} = \widehat{\boldsymbol{\beta}}_{\text{ols}} - \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}_{\text{ols}} - \boldsymbol{c}\right) = \widehat{\boldsymbol{\beta}}_{\text{emd}},$$

the efficient minimum distance estimator. Thus for linear constraints on the linear projection model, efficient GMM equals efficient minimum distance. Thus one convenient method to implement efficient minimum distance is by using GMM methods.

12.16 Distance Test

As in Section 12.13 consider testing the hypothesis $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ where $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta})$ for a given function $\boldsymbol{r}(\boldsymbol{\beta}) : \mathbb{R}^k \to \Theta \subset \mathbb{R}^q$. When $\boldsymbol{r}(\boldsymbol{\beta})$ is non-linear, a better approach than the Wald statistic is use a criterion-based statistic. This is sometimes called the GMM Distance statistic and sometimes called a LR-like statistic (the LR is for likelihood-ratio). The idea was first put forward by Newey and West (1987).

The idea is to compare the unrestricted and restricted estimators by contrasting the criterion functions. The unrestricted estimator takes the form

$$\widehat{\boldsymbol{\beta}}_{\text{gmm}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, J(\boldsymbol{\beta})$$

where

$$\widehat{J}(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \widehat{\boldsymbol{\Omega}}^{-1} \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$$

is the unrestricted GMM criterion which depends on an efficient weight matrix estimate $\widehat{\boldsymbol{\Omega}}$. The minimized value of the criterion is

$$\widehat{J} = \widehat{J}(\widehat{\boldsymbol{\beta}}_{\text{gmm}}).$$

As in Section 12.14, the estimator subject to $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ is

$$\widehat{\boldsymbol{\beta}}_{\text{cgmm}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})=\boldsymbol{\theta}_0}{\operatorname{argmin}}\, \widetilde{J}(\boldsymbol{\beta})$$

where

$$\widetilde{J}(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \widetilde{\boldsymbol{\Omega}}^{-1} \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$$

which depends on an efficient weight matrix estimate $\widetilde{\boldsymbol{\Omega}}$. One possibility is to set $\widetilde{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}$. The minimized value of the criterion is

$$\widetilde{J} = \widetilde{J}(\widehat{\boldsymbol{\beta}}_{\text{cgmm}}).$$

The GMM distance (or LR-like) statistic is the difference in the criterions

$$D = \widetilde{J} - \widehat{J}.$$

The distance test shares the useful feature of LR tests in that it is a natural by-product of the computation of alternative models.

The test has the following large sample distribution.

Theorem 12.16.1 *Under Assumptions 11.14.1 and 8.14.1,* $\boldsymbol{\Omega} > 0$*, and* $\mathbb{H}_0$ *holds, then as* $n \to \infty$*,*

$$D \xrightarrow{d} \chi^2_q.$$

For c *satisfying* $\alpha = 1 - G_q(c)$*,*

$$\Pr(D > c \mid \mathbb{H}_0) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ *if* $D > c$*" has asymptotic size* α*.*

The proof is given in Section 12.24.

Theorem 12.16.1 shows that the distance statistic has a large sample distribution similar to that of Wald and likelihood ratio statistics, and can be interpreted in much the same say. Small values of D mean that imposing the restriction does not result in a large value of the moment equations. Hence the restrictions appear to be compatible with the data. On the other hand, large values

of D mean that imposing the restriction results in a much larger value of the moment equations, implying that the restrictions do not appear to be compatible with the data. The finding that the asymptotic distribution is chi-squared means that it is simple to obtain asymptotic critical values and p-values for the test.

We now discuss the choice of weight matrix. As mentioned above, one simple choice is to set $\widetilde{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}$. In this case we have the following result.

Theorem 12.16.2 *If $\widetilde{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}$ then $D \geq 0$. Furthermore, if $\boldsymbol{r}$ is linear in $\boldsymbol{\beta}$, then D equals the Wald statistic.*

The statement that $\widetilde{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}$ implies $D \geq 0$ follows from the fact that in this case the criterion functions $\widehat{J}(\boldsymbol{\beta}) = \widetilde{J}(\boldsymbol{\beta})$ are identical, so the constrained minimum cannot be smaller than the unconstrained. The statement that linear hypotheses and an efficient weight matrix implies $D = W$ follows from applying the expression for the constrained GMM estimator (12.16) and using the variance matrix formula (12.10).

This result shows some advantages to using the same weight matrix to estimate both $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ and $\widehat{\boldsymbol{\beta}}_{\text{cgmm}}$. In particular, the non-negativity finding motivated Newey and West (1987) to recommend using $\widetilde{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}}$. However, this is not an important advantage. Alternatively, we can set $\widetilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \boldsymbol{z}_i' \widetilde{e}_i^2$ where $\widetilde{e}_i$ are residuals using the constrained estimator. This seems rather natural as in this case $\widehat{J}$ and $\widetilde{J}$ are simple outputs from iterated gmm. In the event that $D < 0$ the test simply fails to reject $\mathbb{H}_0$ at any significance level.

As discussed in Section 9.17, for tests of nonlinear hypotheses the Wald statistic can work quite poorly. In particular, the Wald statistic is affected by how the hypothesis $\boldsymbol{r}(\boldsymbol{\beta})$ is formulated. In contrast, the distance statistic D is not affected by the algebraic formulation of the hypothesis. Current evidence suggests that the D statistic appears to have good sampling properties, and is a preferred test statistic relative to the Wald statistic for nonlinear hypotheses.

In Stata, the command `estat overid` after `ivregress gmm` can be used to report the value of the GMM criterion J. By estimating the two nested GMM regressions the values $\widehat{J}$ and $\widetilde{J}$ can be obtained and D computed.

12.17 Continuously-Updated GMM

An alternative to the two-step GMM estimator can be constructed by letting the weight matrix be an explicit function of $\boldsymbol{\beta}$. These leads to the criterion function

$$J(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta}) \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta})' \right)^{-1} \overline{\boldsymbol{g}}_n(\boldsymbol{\beta}).$$

The $\widehat{\boldsymbol{\beta}}$ which minimizes this function is called the **continuously-updated GMM (CU-GMM)estimator**, and was introduced by L. Hansen, Heaton and Yaron (1996).

A complication is that the continuously-updated criterion $J(\boldsymbol{\beta})$ is not quadratic in $\boldsymbol{\beta}$. This means that minimization requires numerical methods. It may appear that the CU-GMM estimator is the same as the iterated GMM estimator, but this is not the case at all. They solve distinct first-order conditions, and can be quite different in applications.

Relative to traditional GMM, the CU-GMM estimator has lower bias but thicker distributional tails. While it has received considerable theoretical attention, it is not used commonly in applications.

12.18 OverIdentification Test

In Section 11.27 we introduced the Sargan (1958) overidentification test for the 2SLS estimator under the assumption of homoskedasticity. L. Hansen (1982) generalized the test to cover the GMM estimator allowing for general heteroskedasticity.

Recall, overidentified models ($\ell > k$) are special in the sense that there may not be a parameter value $\boldsymbol{\beta}$ such that the moment condition

$$\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\right) = \mathbf{0}$$

holds. Thus the model – the overidentifying restrictions – are testable.

For example, take the linear model $y_i = \boldsymbol{\beta}_1'\boldsymbol{x}_{1i} + \boldsymbol{\beta}_2'\boldsymbol{x}_{2i} + e_i$ with $\mathbb{E}\left(\boldsymbol{x}_{1i}e_i\right) = \mathbf{0}$ and $\mathbb{E}\left(\boldsymbol{x}_{2i}e_i\right) = \mathbf{0}$. It is possible that $\boldsymbol{\beta}_2 = \mathbf{0}$, so that the linear equation may be written as $y_i = \boldsymbol{\beta}_1'\boldsymbol{x}_{1i} + e_i$. However, it is possible that $\boldsymbol{\beta}_2 \neq \mathbf{0}$, and in this case it would be impossible to find a value of $\boldsymbol{\beta}_1$ so that both $\mathbb{E}\left(\boldsymbol{x}_{1i}\left(y_i - \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1\right)\right) = \mathbf{0}$ and $\mathbb{E}\left(\boldsymbol{x}_{2i}\left(y_i - \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1\right)\right) = \mathbf{0}$ hold simultaneously. In this sense an exclusion restriction can be seen as an overidentifying restriction.

Note that $\overline{\boldsymbol{g}}_n \xrightarrow{p} \mathbb{E}\left(\boldsymbol{g}_i\right)$, and thus $\overline{\boldsymbol{g}}_n$ can be used to assess whether or not the hypothesis that $\mathbb{E}\left(\boldsymbol{g}_i\right) = \mathbf{0}$ is true or not. Assuming that an efficient weight matrix estimate is used, the criterion function at the parameter estimates is

$$\begin{aligned} J &= J(\widehat{\boldsymbol{\beta}}_{\text{gmm}}) \\ &= n\ \overline{\boldsymbol{g}}_n'\widehat{\boldsymbol{\Omega}}^{-1}\overline{\boldsymbol{g}}_n \end{aligned}$$

is a quadratic form in $\overline{\boldsymbol{g}}_n$, and is thus a natural test statistic for $\mathbb{H}_0 : \mathbb{E}\left(\boldsymbol{g}_i\right) = \mathbf{0}$. Note that we assume that the criterion function is constructed with an efficient weight matrix estimate. This is important for the distribution theory.

Theorem 12.18.1 *Under Assumption 11.14.1 and $\boldsymbol{\Omega} > 0$, then as $n \to \infty$,*

$$J = J(\widehat{\boldsymbol{\beta}}_{\text{gmm}}) \xrightarrow{d} \chi^2_{\ell-k}.$$

For c satisfying $\alpha = 1 - G_{\ell-k}(c)$,

$$\Pr\left(J > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ if $J > c$" has asymptotic size α.

The proof of the theorem is left to Exercise 12.8.

The degrees of freedom of the asymptotic distribution are the number of overidentifying restrictions. If the statistic J exceeds the chi-square critical value, we can reject the model. Based on this information alone it is unclear what is wrong, but it is typically cause for concern. The GMM overidentification test is a very useful by-product of the GMM methodology, and it is advisable to report the statistic J whenever GMM is the estimation method. When over-identified models are estimated by GMM, it is customary to report the J statistic as a general test of model adequacy.

In Stata, the command `estat overid` afer `ivregress gmm` can be used to implement the overidentification test. The GMM criterion J and its asymptotic p-value using the $\chi^2_{\ell-k}$ distribution are reported.

12.19 Subset OverIdentification Tests

In Section 11.28 we introduced subset overidentification tests for the 2SLS estimator under the assumption of homoskedasticity. In this section we describe how to construct analogous tests for the GMM estimator under general heteroskedasticity.

Recall, subset overidentification tests are used when it is desired to focus attention on a subset of instruments whose validity is questioned. Partition $\boldsymbol{z}_i = (\boldsymbol{z}_{ai}, \boldsymbol{z}_{bi})$ with dimensions ℓ_a and ℓ_b, respectively, where $\boldsymbol{z}_{ai}$ contains the instruments which are believed to be uncorrelated with e_i, and $\boldsymbol{z}_{bi}$ contains the instruments which may be correlated with e_i. It is necessary to select this partition so that $\ell_a > k$, so that the instruments $\boldsymbol{z}_{ai}$ alone identify the parameters. The instruments $\boldsymbol{z}_{bi}$ are potentially valid additional instruments.

Given this partition, the maintained hypothesis is that $\mathbb{E}(\boldsymbol{z}_{ai} e_i) = \mathbf{0}$. The null and alternative hypotheses are

$$\mathbb{H}_0 : \mathbb{E}(\boldsymbol{z}_{bi} e_i) = \mathbf{0}$$
$$\mathbb{H}_1 : \mathbb{E}(\boldsymbol{z}_{bi} e_i) \neq \mathbf{0}.$$

The GMM test is constructed as follows. First, estimate the model by efficient GMM with only the smaller set $\boldsymbol{z}_{ai}$ of instruments. Let $\widetilde{J}$ denote the resulting GMM criterion. Second, estimate the model by efficient GMM with the full set $\boldsymbol{z}_i = (\boldsymbol{z}_{ai}, \boldsymbol{z}_{bi})$ of instruments. Let $\widehat{J}$ denote the resulting GMM criterion. The test statistic is the difference in the criterion functions:

$$C = \widehat{J} - \widetilde{J}.$$

This is similar in form to the GMM distance statistic presented in Section 12.16. The difference is that the distance statistic compares models which differ based on the parameter restrictions, while the C statistic compares models based on different instrument sets.

Typically, the model with the greater instrument set will produce a larger value for J so that $C \geq 0$. However negative values can algebraically occur. That is okay for this simply leads to a non-rejection of $\mathbb{H}_0$.

If the smaller instrument set $\boldsymbol{z}_{ai}$ is just-identified so that $\ell_a = k$ then $\widetilde{J} = 0$ so $C = \widehat{J}$ is simply the standard overidentification test. This is why we have restricted attention to the case $\ell_a > k$.

The test has the following large sample distribution.

Theorem 12.19.1 *Under Assumption 11.14.1, $\boldsymbol{\Omega} > 0$, and $\mathbb{E}(\boldsymbol{z}_{ai}\boldsymbol{x}_i')$ has full rank k, then as $n \to \infty$,*

$$C \xrightarrow{d} \chi^2_{\ell_b}.$$

For c satisfying $\alpha = 1 - G_{\ell_b}(c)$,

$$\Pr(C > c \mid \mathbb{H}_0) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ if $C > c$" has asymptotic size α.

The proof of Theorem 12.19.1 is presented in Section 12.24.

In Stata, the command `estat overid zb` afer `ivregress gmm` can be used to implement a subset overidentification test, where `zb` is the name(s) of the instruments(s) tested for validity. The statistic C and its asymptotic p-value using the $\chi^2_{\ell_2}$ distribution are reported.

12.20 Endogeneity Test

In Section 11.25 we introduced tests for endogeneity in the context of 2SLS estimation. Endogeneity tests are simple to implement in the GMM framework as a subset overidentification test. The model is

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i$$

where the maintained assumption is that the regressors $\boldsymbol{x}_{1i}$ and excluded instruments $\boldsymbol{z}_{2i}$ are exogenous so that $\mathbb{E}(\boldsymbol{x}_{1i}e_i) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{z}_{2i}e_i) = \mathbf{0}$. The question is whether or not $\boldsymbol{x}_{2i}$ is endogenous. Thus the null hypothesis is

$$\mathbb{H}_0 : \mathbb{E}(\boldsymbol{x}_{2i}e_i) = \mathbf{0}$$

with the alternative

$$\mathbb{H}_1 : \mathbb{E}(\boldsymbol{x}_{2i}e_i) \neq \mathbf{0}.$$

The GMM test is constructed as follows. First, estimate the model by efficient GMM using $(\boldsymbol{x}_{1i}, \boldsymbol{z}_{2i})$ as instruments for $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})$. Let $\widetilde{J}$ denote the resulting GMM criterion. Second, estimate the model by efficient GMM using $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i})$ as instruments for $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i})$. Let $\widehat{J}$ denote the resulting GMM criterion. The test statistic is the difference in the criterion functions:

$$C = \widehat{J} - \widetilde{J}.$$

The distribution theory for the test is a special case of the theory of overidentification testing.

Theorem 12.20.1 *Under Assumption 11.14.1, $\boldsymbol{\Omega} > 0$, and $\mathbb{E}(\boldsymbol{z}_{2i}\boldsymbol{x}_{2i}')$ has full rank k_2, then as $n \to \infty$,*

$$C \xrightarrow{d} \chi^2_{k_2}.$$

For c satisfying $\alpha = 1 - G_{\ell_2}(c)$,

$$\Pr(C > c \mid \mathbb{H}_0) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ if $C > c$" has asymptotic size α.

In Stata, the command `estat endogenous` afer `ivregress gmm` can be used to implement the test for endogeneity. The statistic C and its asymptotic p-value using the $\chi^2_{k_2}$ distribution are reported.

12.21 Subset Endogeneity Test

In Section 11.26 we introduced subset endogeneity tests for 2SLS estimation. GMM tests are simple to implement as subset overidentification tests. The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + \boldsymbol{x}_{3i}'\boldsymbol{\beta}_3 + e_i \\ \mathbb{E}(\boldsymbol{z}_i e_i) &= \mathbf{0} \end{aligned}$$

where the instrument vector is $\boldsymbol{z}_i = (\boldsymbol{x}_{1i}, \boldsymbol{z}_{2i})$. The $k_3 \times 1$ variables $\boldsymbol{x}_{3i}$ are treated as endogenous, and the $k_2 \times 1$ variables $\boldsymbol{x}_{2i}$ are treated as potentially endogenous. The hypothesis to test is that $\boldsymbol{x}_{2i}$ is exogenous, or

$$\mathbb{H}_0 : \mathbb{E}(\boldsymbol{x}_{2i}e_i) = \mathbf{0}$$

against

$$\mathbb{H}_1 : \mathbb{E}(\boldsymbol{x}_{2i} e_i) \neq \mathbf{0}.$$

The test requires that $\ell_2 \geq (k_2 + k_3)$ so that the model can be estimated under $\mathbb{H}_1$.

The GMM test is constructed as follows. First, estimate the model by efficient GMM using $(\boldsymbol{x}_{1i}, \boldsymbol{z}_{2i})$ as instruments for $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i})$. Let $\widetilde{J}$ denote the resulting GMM criterion. Second, estimate the model by efficient GMM using $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{z}_{2i})$ as instruments for $(\boldsymbol{x}_{1i}, \boldsymbol{x}_{2i}, \boldsymbol{x}_{3i})$. Let $\widehat{J}$ denote the resulting GMM criterion. The test statistic is the difference in the criterion functions:

$$C = \widehat{J} - \widetilde{J}.$$

The distribution theory for the test is a special case of the theory of overidentification testing.

Theorem 12.21.1 *Under Assumption 11.14.1,* $\boldsymbol{\Omega} > 0$, *and* $\mathbb{E}\left(\boldsymbol{z}_{2i}\left(\boldsymbol{x}_{2i}', \boldsymbol{x}_{3i}'\right)\right)$ *has full rank* $k_2 + k_3$, *then as* $n \to \infty$,

$$C \xrightarrow{d} \chi^2_{k_2}.$$

For c *satisfying* $\alpha = 1 - G_{\ell_2}(c)$,

$$\Pr\left(C > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ *if* $C > c$*" has asymptotic size* α.

In Stata, the command `estat endogenous x2` afer `ivregress gmm` can be used to implement the test for endogeneity, where `x2` is the name(s) of the variable(s) tested for endogeneity. The statistic C and its asymptotic p-value using the $\chi^2_{k_2}$ distribution are reported.

12.22 GMM: The General Case

In its most general form, GMM applies whenever an economic or statistical model implies the $\ell \times 1$ moment condition

$$\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\right) = \mathbf{0}.$$

Often, this is all that is known. Identification requires $\ell \geq k = \dim(\boldsymbol{\beta})$. The GMM estimator minimizes

$$J(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \widehat{\boldsymbol{W}} \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$$

for some weight matrix $\widehat{\boldsymbol{W}}$, where

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\beta}).$$

The efficient GMM estimator can be constructed by setting

$$\widehat{\boldsymbol{W}} = \left(\frac{1}{n} \sum_{i=1}^{n} \widehat{\boldsymbol{g}}_i \widehat{\boldsymbol{g}}_i' - \overline{\boldsymbol{g}}_n \overline{\boldsymbol{g}}_n' \right)^{-1},$$

with $\widehat{\boldsymbol{g}}_i = \boldsymbol{g}(\boldsymbol{w}_i, \widetilde{\boldsymbol{\beta}})$ constructed using a preliminary consistent estimator $\widetilde{\boldsymbol{\beta}}$, perhaps obtained by first setting $\widehat{\boldsymbol{W}} = \boldsymbol{I}_\ell$.

As in the case of the linear model, the weight matrix can be iterated until convergence to obtain the iterated GMM estimator.

Proposition 12.22.1 ***Distribution of Nonlinear GMM Estimator***
Under general regularity conditions,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right).$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Q}\right)\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}$$

with

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{g}_i\boldsymbol{g}_i'\right)$$

and

$$\boldsymbol{Q} = \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\boldsymbol{\beta})\right).$$

If the efficient weight matrix is used then

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}.$$

The proof of this result is omitted as it uses more advanced techniques.

The asymptotic covariance matrices can be estimated by sample counterparts of the population matrices. For the case of a general weight matrix,

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{Q}}\right)^{-1}\left(\widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{W}}\widehat{\boldsymbol{Q}}\right)\left(\widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{W}}\widehat{\boldsymbol{Q}}\right)^{-1}$$

where

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{g}_i(\widehat{\boldsymbol{\beta}}) - \overline{\boldsymbol{g}}\right)\left(\boldsymbol{g}_i(\widehat{\boldsymbol{\beta}}) - \overline{\boldsymbol{g}}\right)'$$

$$\overline{\boldsymbol{g}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{g}_i(\widehat{\boldsymbol{\beta}})$$

and

$$\widehat{\boldsymbol{Q}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\widehat{\boldsymbol{\beta}}).$$

For the case of the iterated efficient weight matrix,

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{Q}}'\widehat{\boldsymbol{\Omega}}^{-1}\widehat{\boldsymbol{Q}}\right)^{-1}.$$

All of the methods discussed in this chapter – Wald tests, constrained estimation, Distance tests, overidentification tests, endogeneity tests – apply similarly to the nonlinear GMM estimator (under the same regularity conditions as the latter).

12.23 Conditional Moment Equation Models

In many contexts, an economic model implies more than an unconditional moment restriction of the form $\mathbb{E}\left(\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\beta})\right) = \mathbf{0}$. It implies a conditional moment restriction of the form

$$\mathbb{E}\left(\boldsymbol{e}_i(\boldsymbol{\beta}) \mid \boldsymbol{z}_i\right) = \mathbf{0}$$

where $\boldsymbol{e}_i(\boldsymbol{\beta})$ is some $s \times 1$ function of the observation and the parameters. In many cases, $s = 1$. The variable $\boldsymbol{z}_i$ is often called an **instrument**.

It turns out that this conditional moment restriction is much more powerful, and restrictive, than the unconditional moment equation model discussed throughout this chapter.

For example, the linear model $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ with instruments $\boldsymbol{z}_i$ falls into this class under the assumption $\mathbb{E}(e_i \mid \boldsymbol{z}_i) = 0$. In this case, $e_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$.

It is also helpful to realize that conventional regression models also fall into this class, except that in this case $\boldsymbol{x}_i = \boldsymbol{z}_i$. For example, in linear regression, $e_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, while in a nonlinear regression model $e_i(\boldsymbol{\beta}) = y_i - \boldsymbol{g}(\boldsymbol{x}_i, \boldsymbol{\beta})$. In a joint model of the conditional mean $\mathbb{E}(y \mid \boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$ and variance $\operatorname{var}(y \mid \boldsymbol{x}) = f(\boldsymbol{x})'\boldsymbol{\gamma}$, then

$$\boldsymbol{e}_i(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{cases} y_i - \boldsymbol{x}_i'\boldsymbol{\beta} \\ \\ (y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 - f(\boldsymbol{x}_i)'\boldsymbol{\gamma} \end{cases}.$$

Here $s = 2$.

Given a conditional moment restriction, an unconditional moment restriction can always be constructed. That is for any $\ell \times 1$ function $\boldsymbol{\phi}(\boldsymbol{z}, \boldsymbol{\beta})$, we can set $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{\phi}(\boldsymbol{z}_i, \boldsymbol{\beta})\, e_i(\boldsymbol{\beta})$ which satisfies $\mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\beta})) = \boldsymbol{0}$ and hence defines an unconditional moment equation model. The obvious problem is that the class of functions $\boldsymbol{\phi}$ is infinite. Which should be selected?

This is equivalent to the problem of selection of the best instruments. If $z_i \in \mathbb{R}$ is a valid instrument satisfying $\mathbb{E}(e_i \mid z_i) = 0$, then z_i, z_i^2, z_i^3, ..., etc., are all valid instruments. Which should be used?

One solution is to construct an infinite list of potent instruments, and then use the first k instruments. How is k to be determined? This is an area of theory still under development. A recent study of this problem is Donald and Newey (2001).

Another approach is to construct the **optimal instrument**. The form was uncovered by Chamberlain (1987). Take the case $s = 1$. Let

$$\boldsymbol{R}_i = \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\beta}} e_i(\boldsymbol{\beta}) \mid \boldsymbol{z}_i\right)$$

and

$$\sigma_i^2 = \mathbb{E}\left(e_i(\boldsymbol{\beta})^2 \mid \boldsymbol{z}_i\right).$$

Then the "optimal instrument" is

$$\boldsymbol{A}_i = -\sigma_i^{-2}\boldsymbol{R}_i$$

so the optimal moment is

$$\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{A}_i e_i(\boldsymbol{\beta}).$$

Setting $\boldsymbol{g}_i(\boldsymbol{\beta})$ to be this choice (which is $k \times 1$, so is just-identified) yields the best GMM estimator possible.

In practice, $\boldsymbol{A}_i$ is unknown, but its form does help us think about construction of optimal instruments.

In the linear model $e_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$, note that

$$\boldsymbol{R}_i = -\mathbb{E}(\boldsymbol{x}_i \mid \boldsymbol{z}_i)$$

and

$$\sigma_i^2 = \mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right),$$

so

$$\boldsymbol{A}_i = \sigma_i^{-2}\mathbb{E}(\boldsymbol{x}_i \mid \boldsymbol{z}_i).$$

In the case of linear regression, $\boldsymbol{x}_i = \boldsymbol{z}_i$, so $\boldsymbol{A}_i = \sigma_i^{-2}\boldsymbol{z}_i$. Hence efficient GMM is equivalently to optimal GLS.

In the case of endogenous variables, note that the efficient instrument $\boldsymbol{A}_i$ involves the estimation of the conditional mean of $\boldsymbol{x}_i$ given $\boldsymbol{z}_i$. In other words, to get the best instrument for $\boldsymbol{x}_i$, we need the best conditional mean model for $\boldsymbol{x}_i$ given $\boldsymbol{z}_i$, not just an arbitrary linear projection. The efficient instrument is also inversely proportional to the conditional variance of e_i. This is the same as the GLS estimator; namely that improved efficiency can be obtained if the observations are weighted inversely to the conditional variance of the errors.

12.24 Technical Proofs*

Proof of Theorem 12.16.1. Set

$$\begin{aligned}\widetilde{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{cgmm}} \\ \widehat{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\text{gmm}}\end{aligned}$$

By standard covariance matrix analysis $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$ and $\widetilde{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$. Thus we can replace $\widehat{\boldsymbol{\Omega}}$ and $\widetilde{\boldsymbol{\Omega}}$ in the criteria without affecting the asymptotic distribution. With this substitution $\widehat{J}(\boldsymbol{\beta}) = \widetilde{J}(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$. From (12.18) and setting $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{cgmm}} - \boldsymbol{\beta}\right) = \left(\boldsymbol{I}_k - \boldsymbol{V}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right) + o_p(1).$$

Thus

$$\begin{aligned}\sqrt{n}\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}_{\text{cgmm}}) &= \frac{1}{\sqrt{n}}\boldsymbol{Z}'\widetilde{\boldsymbol{e}} \\ &= \frac{1}{\sqrt{n}}\boldsymbol{Z}'\widehat{\boldsymbol{e}} + \frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\,\boldsymbol{V}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right) + o_p(1).\end{aligned}$$

The first-order condition for $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ is $\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Omega}^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{e}} = 0$ so the two components in this last expression are orthogonal with respect to the weight matrix $\boldsymbol{\Omega}^{-1}$. Hence

$$\begin{aligned}\widehat{J}(\widehat{\boldsymbol{\beta}}_{\text{cgmm}}) &= \left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\widetilde{\boldsymbol{e}}\right)'\boldsymbol{\Omega}^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\widetilde{\boldsymbol{e}}\right) \\ &= \left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\widehat{\boldsymbol{e}}\right)\boldsymbol{\Omega}^{-1}\left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\widehat{\boldsymbol{e}}\right) \\ &+ n\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right)'\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Omega}^{-1}\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\,\boldsymbol{V}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'n\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right) \\ &+ o_p(1) \\ &= \widehat{J}(\widehat{\boldsymbol{\beta}}_{\text{gmm}}) + n\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right)'\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right) + o_p(1).\end{aligned}$$

Thus

$$\begin{aligned}D &= \widehat{J}(\widehat{\boldsymbol{\beta}}_{\text{cgmm}}) - \widehat{J}(\widehat{\boldsymbol{\beta}}_{\text{gmm}}) \\ &= n\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right)'\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\widehat{\boldsymbol{\beta}}_{\text{gmm}} - \boldsymbol{\beta}\right) + o_p(1)\end{aligned}$$

which converges in distribution to χ^2_q as claimed. ■

Proof of Theorem 12.19.1. Let $\widetilde{\boldsymbol{\beta}}$ denote the GMM estimate obtained with the instrument set

$\boldsymbol{z}_{ai}$ and let $\widehat{\boldsymbol{\beta}}$ denote the GMM estimates obtained with the instrument set $\boldsymbol{z}_i$. Set

$$\widetilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$$
$$\widehat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$
$$\widetilde{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{z}_{ai}\boldsymbol{z}_{ai}'\widetilde{e}_i^2$$
$$\widehat{\boldsymbol{\Omega}} = n^{-1}\sum_{i=1}^{n} \boldsymbol{z}_i\boldsymbol{z}_i'\widehat{e}_i^2$$

Let $\boldsymbol{R}$ be the $\ell \times \ell_a$ selector matrix so that $\boldsymbol{z}_{ai} = \boldsymbol{R}'\boldsymbol{z}_i$. Note that

$$\widetilde{\boldsymbol{\Omega}} = \boldsymbol{R}' n^{-1}\sum_{i=1}^{n} \boldsymbol{z}_i\boldsymbol{z}_i'\widetilde{e}_i^2\boldsymbol{R}.$$

By standard covariance matrix analysis, $\widehat{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{\Omega}$ and $\widetilde{\boldsymbol{\Omega}} \xrightarrow{p} \boldsymbol{R}'\boldsymbol{\Omega}\boldsymbol{R}$. Also, $\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{Q}$, say. By the CLT, $n^{-1/2}\boldsymbol{Z}'\boldsymbol{e} \xrightarrow{d} \mathrm{Z}$ where $\mathrm{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega})$. Then

$$n^{-1/2}\boldsymbol{Z}'\widehat{\boldsymbol{e}} = \left(\boldsymbol{I}_\ell - \left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\widehat{\boldsymbol{\Omega}}^{-1}\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\widehat{\boldsymbol{\Omega}}^{-1}\right) n^{-1/2}\boldsymbol{Z}'\boldsymbol{e}$$
$$\xrightarrow{d} \left(\boldsymbol{I}_\ell - \boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\right)\mathrm{Z}$$

and

$$n^{-1/2}\boldsymbol{Z}_a'\widetilde{\boldsymbol{e}} = \boldsymbol{R}'\left(\boldsymbol{I}_\ell - \left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{R}\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{R}'\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{R}\widetilde{\boldsymbol{\Omega}}^{-1}\boldsymbol{R}'\right) n^{-1/2}\boldsymbol{Z}'\boldsymbol{e}$$
$$\xrightarrow{d} \boldsymbol{R}'\left(\boldsymbol{I}_\ell - \boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{\Omega}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{\Omega}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\mathrm{Z}$$

jointly. Thus

$$\widehat{J} \xrightarrow{d} \mathrm{Z}'\left(\boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\right)\mathrm{Z}$$

and

$$\widetilde{J} \xrightarrow{d} \mathrm{Z}'\left(\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{\Omega}\boldsymbol{R}\right)^{-1}\boldsymbol{R}' - \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{\Omega}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{\Omega}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{\Omega}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\mathrm{Z}.$$

By linear rotations of Z and $\boldsymbol{R}$ we can set $\boldsymbol{\Omega} = \boldsymbol{I}_\ell$ to simplify the notation. It follows that

$$C \xrightarrow{d} \mathrm{Z}'\boldsymbol{A}\mathrm{Z}$$

where

$$\boldsymbol{A} = \left(\boldsymbol{I}_\ell - \boldsymbol{P}_{\boldsymbol{Q}} - \boldsymbol{P}_{\boldsymbol{R}} + \boldsymbol{P}_{\boldsymbol{R}}\boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{P}_{\boldsymbol{R}}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{P}_{\boldsymbol{R}}\right),$$

$\boldsymbol{P}_{\boldsymbol{R}} = \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{R}\right)^{-1}\boldsymbol{R}'$, $\boldsymbol{P}_{\boldsymbol{Q}} = \boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'$, and $\mathrm{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_\ell)$. This is a quadratic form in a standard normal vector, and the matrix $\boldsymbol{A}$ is idempotent (this is straightforward to check). It is thus distributed as χ^2_d with degrees of freedom d equal to the rank of $\boldsymbol{A}$. This is

$$\begin{aligned}\operatorname{rank}(\boldsymbol{A}) &= \operatorname{tr}\left(\boldsymbol{I}_\ell - \boldsymbol{P}_{\boldsymbol{Q}} - \boldsymbol{P}_{\boldsymbol{R}} + \boldsymbol{P}_{\boldsymbol{R}}\boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{P}_{\boldsymbol{R}}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{P}_{\boldsymbol{R}}\right) \\ &= \ell - k - \ell_a + k \\ &= \ell_b.\end{aligned}$$

Thus the asymptotic distribution of C is $\chi^2_{\ell_b}$ as claimed. ■

Exercises

Exercise 12.1 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \\ e_i^2 &= \boldsymbol{z}_i'\boldsymbol{\gamma} + \eta_i \\ \mathbb{E}\left(\boldsymbol{z}_i \eta_i\right) &= \mathbf{0}. \end{aligned}$$

Find the method of moments estimators $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\gamma}}\right)$ for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

Exercise 12.2 Take the single equation

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{Z}\right) &= \mathbf{0} \end{aligned}$$

Assume $\mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right) = \sigma^2$. Show that if $\widehat{\boldsymbol{\beta}}_{\text{gmm}}$ is the GMM estimated by GMM with weight matrix $\boldsymbol{W}_n = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$, then

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \sigma^2\left(\boldsymbol{Q}'\boldsymbol{M}^{-1}\boldsymbol{Q}\right)^{-1}\right)$$

where $\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i'\right)$ and $\boldsymbol{M} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)$.

Exercise 12.3 Take the model $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ with $\mathbb{E}\left(\boldsymbol{z}_i e_i\right) = \mathbf{0}$. Let $\widetilde{e}_i = y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}$ where $\widetilde{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ (e.g. a GMM estimator with arbitrary weight matrix). Define an estimate of the optimal GMM weight matrix

$$\widehat{\boldsymbol{W}} = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i\boldsymbol{z}_i'\widetilde{e}_i^2\right)^{-1}.$$

Show that $\widehat{\boldsymbol{W}} \xrightarrow{p} \boldsymbol{\Omega}^{-1}$ where $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)$.

Exercise 12.4 In the linear model estimated by GMM with general weight matrix $\boldsymbol{W}$, the asymptotic variance of $\widehat{\boldsymbol{\beta}}_{GMM}$ is

$$\boldsymbol{V} = \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Q}\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}$$

(a) Let $\boldsymbol{V}_0$ be this matrix when $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}$. Show that $\boldsymbol{V}_0 = \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}$.

(b) We want to show that for any $\boldsymbol{W}$, $\boldsymbol{V} - \boldsymbol{V}_0$ is positive semi-definite (for then $\boldsymbol{V}_0$ is the smaller possible covariance matrix and $\boldsymbol{W} = \boldsymbol{\Omega}^{-1}$ is the efficient weight matrix). To do this, start by finding matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that $\boldsymbol{V} = \boldsymbol{A}'\boldsymbol{\Omega}\boldsymbol{A}$ and $\boldsymbol{V}_0 = \boldsymbol{B}'\boldsymbol{\Omega}\boldsymbol{B}$.

(c) Show that $\boldsymbol{B}'\boldsymbol{\Omega}\boldsymbol{A} = \boldsymbol{B}'\boldsymbol{\Omega}\boldsymbol{B}$ and therefore that $\boldsymbol{B}'\boldsymbol{\Omega}\left(\boldsymbol{A} - \boldsymbol{B}\right) = \mathbf{0}$.

(d) Use the expressions $\boldsymbol{V} = \boldsymbol{A}'\boldsymbol{\Omega}\boldsymbol{A}$, $\boldsymbol{A} = \boldsymbol{B} + \left(\boldsymbol{A} - \boldsymbol{B}\right)$, and $\boldsymbol{B}'\boldsymbol{\Omega}\left(\boldsymbol{A} - \boldsymbol{B}\right) = \mathbf{0}$ to show that $\boldsymbol{V} \geq \boldsymbol{V}_0$.

Exercise 12.5 The equation of interest is

$$\begin{aligned} y_i &= \boldsymbol{m}(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i \\ \mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= \mathbf{0}. \end{aligned}$$

The observed data is $(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$. $\boldsymbol{z}_i$ is $\ell \times 1$ and $\boldsymbol{\beta}$ is $k \times 1$, $\ell \geq k$. Show how to construct an efficient GMM estimator for $\boldsymbol{\beta}$.

Exercise 12.6 As a continuation of Exercise 11.7, derive the efficient GMM estimator using the instrument $\boldsymbol{z}_i = (x_i \quad x_i^2)'$. Does this differ from 2SLS and/or OLS?

Exercise 12.7 In the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ with $\mathbb{E}(\boldsymbol{x}_i e_i) = \boldsymbol{0}$, a Generalized Method of Moments (GMM) criterion function for $\boldsymbol{\beta}$ is defined as

$$J(\boldsymbol{\beta}) = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \boldsymbol{X} \widehat{\boldsymbol{\Omega}}^{-1} \boldsymbol{X}' (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{12.19}$$

where $\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i' \widehat{e}_i^2$, $\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ are the OLS residuals, and $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ is LS. The GMM estimator of $\boldsymbol{\beta}$, subject to the restriction $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{0}$, is defined as

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})=\boldsymbol{0}}{\operatorname{argmin}} J_n(\boldsymbol{\beta}).$$

The GMM test statistic (the distance statistic) of the hypothesis $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{0}$ is

$$D = J(\widetilde{\boldsymbol{\beta}}) = \min_{\boldsymbol{r}(\boldsymbol{\beta})=\boldsymbol{0}} J(\boldsymbol{\beta}). \tag{12.20}$$

(a) Show that you can rewrite $J(\boldsymbol{\beta})$ in (12.19) as

$$J(\boldsymbol{\beta}) = n\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1} \left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)$$

thus $\widetilde{\boldsymbol{\beta}}$ is the same as the minimum distance estimator.

(b) Show that under linear hypotheses the distance statistic D in (12.20) equals the Wald statistic.

Exercise 12.8 Take the linear model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{z}_i e_i) &= \boldsymbol{0}. \end{aligned}$$

and consider the GMM estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Let

$$J = n\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Omega}}^{-1}\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})$$

denote the test of overidentifying restrictions. Show that $J \xrightarrow{d} \chi^2_{\ell-k}$ as $n \to \infty$ by demonstrating each of the following:

(a) Since $\boldsymbol{\Omega} > 0$, we can write $\boldsymbol{\Omega}^{-1} = \boldsymbol{C}\boldsymbol{C}'$ and $\boldsymbol{\Omega} = \boldsymbol{C}'^{-1}\boldsymbol{C}^{-1}$

(b) $J = n\left(\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})\right)' \left(\boldsymbol{C}'\widehat{\boldsymbol{\Omega}}\boldsymbol{C}\right)^{-1} \boldsymbol{C}'\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})$

(c) $\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}) = \boldsymbol{D}_n\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$ where

$$\boldsymbol{D}_n = \boldsymbol{I}_\ell - \boldsymbol{C}'\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\widehat{\boldsymbol{\Omega}}^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\widehat{\boldsymbol{\Omega}}^{-1}\boldsymbol{C}'^{-1}$$

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{e}.$$

(d) $\boldsymbol{D}_n \xrightarrow{p} \boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'$ where $\boldsymbol{R} = \boldsymbol{C}'\mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i')$

(e) $n^{1/2}\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) \xrightarrow{d} \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_\ell)$

(f) $J \xrightarrow{d} \boldsymbol{u}' \left(\boldsymbol{I}_\ell - \boldsymbol{R} \left(\boldsymbol{R}'\boldsymbol{R} \right)^{-1} \boldsymbol{R}' \right) \boldsymbol{u}$

(g) $\boldsymbol{u}' \left(\boldsymbol{I}_\ell - \boldsymbol{R} \left(\boldsymbol{R}'\boldsymbol{R} \right)^{-1} \boldsymbol{R}' \right) \boldsymbol{u} \sim \chi^2_{\ell-k}$.

Hint: $\boldsymbol{I}_\ell - \boldsymbol{R} \left(\boldsymbol{R}'\boldsymbol{R} \right)^{-1} \boldsymbol{R}'$ is a projection matrix.

Exercise 12.9 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= \boldsymbol{0} \end{aligned}$$

y_i scalar, $\boldsymbol{x}_i$ a k vector and $\boldsymbol{z}_i$ an ℓ vector, $\ell \geq k$. Assume iid observations. Consider the statistic

$$\begin{aligned} J_n(\beta) &= n\overline{\boldsymbol{m}}_n(\boldsymbol{\beta})' \boldsymbol{W} \overline{\boldsymbol{m}}_n(\boldsymbol{\beta}) \\ \overline{\boldsymbol{m}}_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta} \right) \end{aligned}$$

for some weight matrix $\boldsymbol{W} > 0$.

(a) Take the hypothesis

$$\mathbb{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$

Derive the asymptotic distribution of $J_n(\boldsymbol{\beta}_0)$ under $\mathbb{H}_0$ as $n \to \infty$.

(b) What choice for $\boldsymbol{W}$ yields a known asymptotic distribution in part (a)? (Be specific about degrees of freedom.)

(c) Write down an appropriate estimator $\widehat{\boldsymbol{W}}$ for $\boldsymbol{W}$ which takes advantage of $\mathbb{H}_0$. (You do not need to demonstrate consistency or unbiasedness.)

(d) Describe an asymptotic test of $\mathbb{H}_0$ against $\mathbb{H}_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ based on this statistic.

(e) Use the result in part (d) to construct a confidence region for $\boldsymbol{\beta}$. What can you say about the form of this region? For example, does the confidence region take the form of an ellipse, similar to conventional confidence regions?

Exercise 12.10 Consider the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= \boldsymbol{0} \qquad (12.21) \\ \boldsymbol{R}'\boldsymbol{\beta} &= \boldsymbol{0} \qquad (12.22) \end{aligned}$$

with y_i scalar, $\boldsymbol{x}_i$ a k vector and $\boldsymbol{z}_i$ an ℓ vector with $\ell > k$. The matrix $\boldsymbol{R}$ is $k \times q$ with $1 \leq q < k$. You have a random sample $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i : i = 1, ..., n)$.

For simplicity, assume the "efficient" weight matrix $\boldsymbol{W} = \left(\mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i' e_i^2 \right) \right)^{-1}$ is known.

(a) Write out the GMM estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ given the moment conditions (12.21) but ignoring constraint (12.22).

(b) Write out the GMM estimator $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ given the moment conditions (12.21) and constraint (12.22).

(c) Find the asymptotic distribution of $\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \right)$ as $n \to \infty$ under the assumption that (12.21) and (12.22) are correct.

Exercise 12.11 The observed data is $\{y_i, x_i, \boldsymbol{z}_i\} \in \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}^\ell$, $k > 1$ and $\ell > k > 1$, $i = 1, ..., n$. The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{z}_i e_i) &= 0 \end{aligned} \tag{12.23}$$

(a) Given a weight matrix $\boldsymbol{W} > 0$, write down the GMM estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.

(b) Suppose the model is misspecified in that

$$\begin{aligned} e_i &= \delta n^{-1/2} + u_i \\ \mathbb{E}(u_i \mid \boldsymbol{z}_i) &= 0 \end{aligned} \tag{12.24}$$

with $\boldsymbol{\mu}_{\boldsymbol{z}} = \mathbb{E}(\boldsymbol{z}_i) \neq 0$ and $\delta \neq 0$. Show that (12.24) implies (12.23) is false.

(c) Express $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ as a function of $\boldsymbol{W}$, n, δ, and the variables $(\boldsymbol{x}_i, \boldsymbol{z}_i, u_i)$.

(d) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ under Assumption (12.24).

Exercise 12.12 The model is

$$\begin{aligned} y_i &= z_i\beta + x_i\gamma + e_i \\ \mathbb{E}(e_i \mid x_i) &= 0 \end{aligned}$$

Thus z_i is potentially endogenous and x_i is exogenous. Assume that z_i and x_i are scalar. Someone suggests estimating (β, γ) by GMM, using the pair (x_i, x_i^2) as the instruments. Is this feasible? Under what conditions, if any, (in additional to those described above) is this a valid estimator?

Exercise 12.13 The observations are iid, $(y_i, \boldsymbol{x}_i, \boldsymbol{q}_i : i = 1, ..., n)$, where $\boldsymbol{x}_i$ is $k \times 1$ and $\boldsymbol{q}_i$ is $m \times 1$. The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= 0 \\ \mathbb{E}(\boldsymbol{q}_i e_i) &= 0 \end{aligned}$$

Find the efficient GMM estimator for $\boldsymbol{\beta}$.

Exercise 12.14 You want to estimate $\mu = \mathbb{E}(y_i)$ under the assumption that $\mathbb{E}(x_i) = 0$, where y_i and x_i are scalar and observed from a random sample. Find an efficient GMM estimator for μ.

Exercise 12.15 Consider the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{z}_i e_i) &= 0 \\ \boldsymbol{R}'\boldsymbol{\beta} &= 0 \end{aligned}$$

The dimensions are $\boldsymbol{x} \in R^k$, $\boldsymbol{z} \in R^\ell$, $\ell > k$. The matrix $\boldsymbol{R}$ is $k \times q$, $1 \leq q < k$. Derive an efficient GMM estimator for $\boldsymbol{\beta}$ for this model.

Exercise 12.16 Take the linear equation $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e$, and consider the following estimators of $\boldsymbol{\beta}$.

1. $\widehat{\boldsymbol{\beta}}$: 2SLS using the instruments $\boldsymbol{z}_{1i}$
2. $\widetilde{\boldsymbol{\beta}}$: 2SLS using the instruments $\boldsymbol{z}_{1i}$

3. $\overline{\boldsymbol{\beta}}$: GMM using the instruments $\boldsymbol{z}_i = (\boldsymbol{z}_{1i}, \boldsymbol{z}_{2i})$ and the weight matrix

$$\boldsymbol{W} = \begin{pmatrix} (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\lambda & 0 \\ 0 & (\boldsymbol{Z}_2'\boldsymbol{Z}_2)^{-1}(1-\lambda) \end{pmatrix}$$

for $\lambda \in (0,1)$. Find an expression for $\overline{\boldsymbol{\beta}}$ which shows that it is a specific weighted average of $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$.

Exercise 12.17 Consider the just-identified model

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= 0 \end{aligned}$$

where $\boldsymbol{x}_i = (\boldsymbol{x}_{1i}' \ \boldsymbol{x}_{2i}')'$ and $\boldsymbol{z}_i$ are $k \times 1$. We want to test $\mathbb{H}_0 : \boldsymbol{\beta}_1 = 0$. Three econometricians are called to advise on how to test $\mathbb{H}_0$.

- Econometrician 1 proposes testing $\mathbb{H}_0$ by a Wald statistic.
- Econometrician 2 suggests testing $\mathbb{H}_0$ by the GMM Distance Statistic.
- Econometrician 3 suggests testing $\mathbb{H}_0$ using the test of overidentifying restrictions.

You are asked to settle this dispute. Explain the advantages and/or disadvantages of the different procedures, in this specific context.

Exercise 12.18 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= \mathbf{0} \\ \boldsymbol{\beta} &= \boldsymbol{Q}\boldsymbol{\theta} \end{aligned}$$

where $\boldsymbol{\beta}$ is $k \times 1$, $\boldsymbol{Q}$ is $k \times m$ with $m < k$, and $\boldsymbol{Q}$ is known. Assume that the observations $(y_i, \boldsymbol{x}_i)$ are i.i.d. across $i = 1, ..., n$.

Under these assumptions, what is the efficient estimator of $\boldsymbol{\theta}$?

Exercise 12.19 Take the model

$$\begin{aligned} y_i &= \theta + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= 0 \end{aligned}$$

with $(y_i, \boldsymbol{x}_i)$ a random sample. y_i is real-valued and $\boldsymbol{x}_i$ is $k \times 1$, $k > 1$.

(a) Find the efficient GMM estimator of θ.

(b) Is this model over-identified or just-identified?

(c) Find the GMM test statistic for over-identification.

Exercise 12.20 Continuation of Exercise 11.23, based on the empirical work reported in Acemoglu, Johnson and Robinson (2001)

(a) Re-estimate the model estimated part (j) by efficient GMM. I suggest that you use the 2SLS estimates as the first-step to get the weight matrix, and then calculate the GMM estimator from this weight matrix without further iteration. Report the estimates and standard errors.

(b) Calculate and report the J statistic for overidentification.

(c) Compare the GMM and 2SLS estimates. Discuss your findings

Exercise 12.21 Continuation of Exercise 11.24, which involved estimation of a wage equation by 2SLS.

(a) Re-estimate the model in part (a) by efficient GMM. Do the results change meaningfully?

(b) Re-estimate the model in part (d) by efficient GMM. Do the results change meaningfully?

(c) Report the J statistic for overidentification.

Chapter 13

The Bootstrap

13.1 Definition of the Bootstrap

Let F denote the distribution function for the population of observations $(y_i, \boldsymbol{x}_i)$. Let

$$T = T\left((y_1, \boldsymbol{x}_1), ..., (y_n, \boldsymbol{x}_n), F\right)$$

be a statistic of interest, for example an estimator $\widehat{\theta}$ or a t-statistic $\left(\widehat{\theta} - \theta\right)/s(\widehat{\theta})$. Note that we write T as possibly a function of F. For example, the t-statistic is a function of the parameter $\theta = \theta(F)$ which itself is a function of F.

The exact CDF of T when the data are sampled from the distribution F is

$$G_n(u, F) = \Pr(T \leq u \mid F)$$

In general, $G_n(u, F)$ depends on F and n, meaning that G changes as F or n changes.

Ideally, inference would be based on $G_n(u, F)$. This is generally impossible since F is unknown.

Asymptotic inference is based on approximating $G_n(u, F)$ with $G(u, F) = \lim_{n\to\infty} G_n(u, F)$. When $G(u, F) = G(u)$ does not depend on F, we say that T is asymptotically pivotal and use the distribution function $G(u)$ for inferential purposes.

In a seminal contribution, Efron (1979) proposed the bootstrap, which makes a different approximation. The unknown F is replaced by a consistent estimate $\widehat{F}$ (one choice is discussed in the next section). Plugged into $G_n(u, F)$ we obtain

$$G_n^*(u) = G_n(u, \widehat{F}). \tag{13.1}$$

We call G_n^* the bootstrap distribution. Bootstrap inference is based on $G_n^*(u)$.

Let $(y_i^*, \boldsymbol{x}_i^*)$ denote random variables from the distribution $\widehat{F}$. A random sample $\{(y_i^*, \boldsymbol{x}_i^*) : i = 1, ..., n\}$ from this distribution is called the **bootstrap data**. The statistic $T^* = T\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), \widehat{F}\right)$ constructed on this sample is a random variable with distribution G_n^*. That is, $\Pr(T^* \leq u) = G_n^*(u)$. We call T^* the **bootstrap statistic**. The distribution of T^* is identical to that of T when the true CDF is $\widehat{F}$ rather than F.

The bootstrap distribution is itself random, as it depends on the sample through the estimator $\widehat{F}$.

In the next sections we describe computation of the bootstrap distribution.

13.2 The Empirical Distribution Function

Recall that $F(y, \boldsymbol{x}) = \Pr\left(y_i \leq y, \boldsymbol{x}_i \leq \boldsymbol{x}\right) = \mathbb{E}\left(1\left(y_i \leq y\right) 1\left(\boldsymbol{x}_i \leq \boldsymbol{x}\right)\right)$, where $1(\cdot)$ is the indicator function. This is a population moment. The method of moments estimator is the corresponding

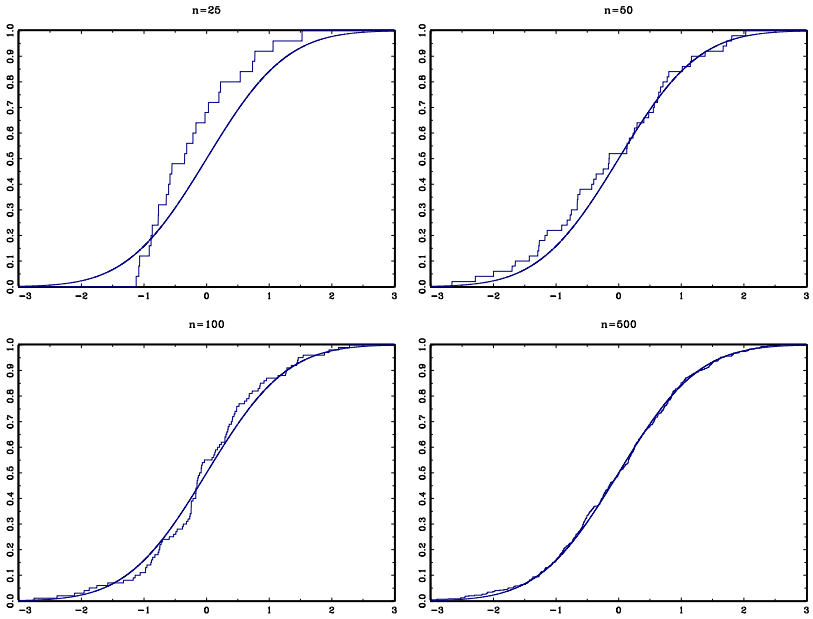


Figure 13.1: Empirical Distribution Functions

sample moment:

$$\widehat{F}(y, \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i \leq y)\, 1(\boldsymbol{x}_i \leq \boldsymbol{x}). \tag{13.2}$$

$\widehat{F}(y, \boldsymbol{x})$ is called the empirical distribution function (EDF) and is a nonparametric estimate of F. Note that while F may be either discrete or continuous, $\widehat{F}$ is by construction a step function.

The EDF is a consistent estimator of the CDF. To see this, note that for any $(y, \boldsymbol{x})$, $1(y_i \leq y)\, 1(\boldsymbol{x}_i \leq \boldsymbol{x})$ is an iid random variable with expectation $F(y, \boldsymbol{x})$. Thus by the WLLN (Theorem 6.4.2), $\widehat{F}(y, \boldsymbol{x}) \xrightarrow{p} F(y, \boldsymbol{x})$. Furthermore, by the CLT (Theorem 6.8.1),

$$\sqrt{n}\left(\widehat{F}(y, \boldsymbol{x}) - F(y, \boldsymbol{x})\right) \xrightarrow{d} \mathrm{N}(0, F(y, \boldsymbol{x})(1 - F(y, \boldsymbol{x}))).$$

To see the effect of sample size on the EDF, in Figure 13.1, I have plotted the EDF and true CDF for three random samples of size $n = 25$, 50, 100, and 500. The random draws are from the $\mathrm{N}(0, 1)$ distribution. For $n = 25$, the EDF is only a crude approximation to the CDF, but the approximation appears to improve for the large n. In general, as the sample size gets larger, the EDF step function gets uniformly close to the true CDF.

The EDF is a valid discrete probability distribution which puts probability mass $1/n$ at each pair $(y_i, \boldsymbol{x}_i)$, $i = 1, ..., n$. Notationally, it is helpful to think of a random pair $(y_i^*, \boldsymbol{x}_i^*)$ with the distribution $\widehat{F}$. That is,

$$\Pr(y_i^* \leq y, \boldsymbol{x}_i^* \leq \boldsymbol{x}) = \widehat{F}(y, \boldsymbol{x}).$$

We can easily calculate the moments of functions of $(y_i^*, \boldsymbol{x}_i^*)$:

$$\begin{aligned}
\mathbb{E}\left(h\left(y_i^*, \boldsymbol{x}_i^*\right)\right) &= \int h(y, \boldsymbol{x}) d\widehat{F}(y, \boldsymbol{x}) \\
&= \sum_{i=1}^{n} h\left(y_i, \boldsymbol{x}_i\right) \Pr\left(y_i^* = y_i, \boldsymbol{x}_i^* = \boldsymbol{x}_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} h\left(y_i, \boldsymbol{x}_i\right),
\end{aligned}$$

the empirical sample average.

13.3 Nonparametric Bootstrap

The **nonparametric bootstrap** is obtained when the bootstrap distribution (13.1) is defined using the EDF (13.2) as the estimate $\widehat{F}$ of F.

Since the EDF $\widehat{F}$ is a multinomial (with n support points), in principle the distribution G_n^* could be calculated by direct methods. However, as there are $\binom{2n-1}{n}$ possible samples $\{(y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*)\}$, such a calculation is computationally infeasible. The popular alternative is to use simulation to approximate the distribution. The algorithm is identical to our discussion of Monte Carlo simulation, with the following points of clarification:

- The sample size n used for the simulation is the same as the sample size.
- The random vectors $(y_i^*, \boldsymbol{x}_i^*)$ are drawn randomly from the empirical distribution. This is equivalent to sampling a pair $(y_i, \boldsymbol{x}_i)$ randomly from the sample.

The bootstrap statistic $T^* = T\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), \widehat{F}\right)$ is calculated for each bootstrap sample. This is repeated B times. B is known as the number of bootstrap replications. A theory for the determination of the number of bootstrap replications B has been developed by Andrews and Buchinsky (2000). It is desirable for B to be large, so long as the computational costs are reasonable. $B = 1000$ typically suffices.

When the statistic T is a function of F, it is typically through dependence on a parameter. For example, the t-ratio $\left(\widehat{\theta} - \theta\right)/s(\widehat{\theta})$ depends on θ. As the bootstrap statistic replaces F with $\widehat{F}$, it similarly replaces θ with $\theta^* = \theta(\widehat{F})$, the value of θ implied by $\widehat{F}$. Typically $\theta^* = \widehat{\theta}$, the parameter estimate. (When in doubt use $\widehat{\theta}$.)

Sampling from the EDF is particularly easy. Since $\widehat{F}$ is a discrete probability distribution putting probability mass $1/n$ at each sample point, sampling from the EDF is equivalent to random sampling a pair $(y_i, \boldsymbol{x}_i)$ from the observed data **with replacement**. In consequence, a bootstrap sample $\{(y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*)\}$ will necessarily have some ties and multiple values, which is generally not a problem.

13.4 Bootstrap Estimation of Bias and Variance

The bias of $\widehat{\theta}$ is $\tau = \mathbb{E}(\widehat{\theta} - \theta)$. The bootstrap counterparts are $\widehat{\theta}^* = \widehat{\theta}((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*))$ and $\tau^* = \mathbb{E}(\widehat{\theta}^* - \theta^*)$. The latter can be estimated by the simulation described in the previous section. This estimator is

$$\begin{aligned}\widehat{\tau}^* &= \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_b^* - \widehat{\theta}\right) \\ &= \overline{\widehat{\theta}^*} - \widehat{\theta}.\end{aligned}$$

If $\widehat{\theta}$ is biased, it might be desirable to construct a biased-corrected estimator for θ (one with reduced bias). Ideally, this would be

$$\widetilde{\theta} = \widehat{\theta} - \tau,$$

but τ is unknown. The (estimated) bootstrap biased-corrected estimator is

$$\begin{aligned}\widetilde{\theta}^* &= \widehat{\theta} - \widehat{\tau}^* \\ &= \widehat{\theta} - (\overline{\widehat{\theta}^*} - \widehat{\theta}) \\ &= 2\widehat{\theta} - \overline{\widehat{\theta}^*}.\end{aligned}$$

Note, in particular, that the biased-corrected estimator is not $\overline{\widehat{\theta}^*}$. Intuitively, the bootstrap makes the following experiment. Suppose that $\widehat{\theta}$ is the truth. Then what is the average value of $\widehat{\theta}$ calculated from such samples? The answer is $\overline{\widehat{\theta}^*}$. If this is lower than $\widehat{\theta}$, this suggests that the estimator is downward-biased, so a biased-corrected estimator of θ should be larger than $\widehat{\theta}$, and the best guess is the difference between $\widehat{\theta}$ and $\overline{\widehat{\theta}^*}$. Similarly if $\overline{\widehat{\theta}^*}$ is higher than $\widehat{\theta}$, then the estimator is upward-biased and the biased-corrected estimator should be lower than $\widehat{\theta}$.

Recall that variance of $\widehat{\theta}$ is

$$V_n = \mathbb{E}\left((\widehat{\theta} - \mathbb{E}\left(\widehat{\theta}\right))^2\right).$$

The bootstrap analog is the variance of $\widehat{\theta}^*$ which is

$$V^* = \mathbb{E}\left((\widehat{\theta}^* - \mathbb{E}\left(\widehat{\theta}^*\right))^2\right).$$

The simulation estimate is

$$\widehat{V}_n^* = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\theta}_b^* - \overline{\widehat{\theta}^*}\right)^2.$$

A bootstrap standard error for $\widehat{\theta}$ is the square root of the bootstrap estimate of variance, $s^*(\widehat{\theta}) = \sqrt{\widehat{V}_n^*}$. These are frequently reported in applied economics instead of asymptotic standard errors.

13.5 Percentile Intervals

Consider an estimator $\widehat{\theta}$ for θ and suppose we wish to construct a confidence interval for θ. Let $G_n(u, F)$ denote the distribution of $\widehat{\theta}$ and let $q(\alpha) = q(\alpha, F)$ denote its quantile function. This is the function which solves

$$G_n(q(\alpha), F) = \alpha.$$

Let $q^*(\alpha) = q(\alpha, \widehat{F})$ denote the quantile function of the bootstrap distribution. Note that this function will change depending on the underlying statistic T whose distribution is G_n.

In $100(1-\alpha)\%$ of samples, $\widehat{\theta}$ lies in the region $[q(\alpha/2), q(1-\alpha/2)]$. This motivates a confidence interval proposed by Efron:

$$\widehat{C}_1 = [q^*(\alpha/2), \quad q^*(1-\alpha/2)].$$

This is often called the **percentile confidence interval**.

Computationally, the quantile $q^*(\alpha)$ is estimated by $\widehat{q}^*(\alpha)$, the α^{th} sample quantile of the simulated statistics $\{T_1^*, ..., T_B^*\}$, as discussed in the section on Monte Carlo simulation. The $1-\alpha$ Efron percentile interval is then $[\widehat{q}^*(\alpha/2), \quad \widehat{q}^*(1-\alpha/2)]$.

The interval $\widehat{C}_1$ is a popular bootstrap confidence interval often used in empirical practice. This is because it is easy to compute, simple to motivate, was popularized by Efron early in the history of the bootstrap, and also has the feature that it is translation invariant. That is, if we define $\phi = f(\theta)$ as the parameter of interest for a monotonically increasing function f, then percentile method applied to this problem will produce the confidence interval $[f(q^*(\alpha/2)), \quad f(q^*(1-\alpha/2))]$, which is a naturally good property.

However, as we show now, $\widehat{C}_1$ can work poorly unless the sampling distribution of $\widehat{\theta}$ is symmetric about θ.

It will be useful if we introduce an alternative definition of $\widehat{C}_1$. Let $q(\alpha)$ and $q^*(\alpha)$ be the quantile functions of $\widehat{\theta} - \theta$ and $\widehat{\theta}^* - \widehat{\theta}$ (These are the original quantiles, with θ and $\widehat{\theta}$ subtracted.) Then $\widehat{C}_1$ can alternatively be written as

$$\widehat{C}_1 = [\widehat{\theta} + q^*(\alpha/2), \quad \hat{\theta} + q^*(1-\alpha/2)].$$

This is a bootstrap estimate of the "ideal" confidence interval

$$\widehat{C}_1^0 = [\widehat{\theta} + q(\alpha/2), \quad \widehat{\theta} + q(1-\alpha/2)].$$

The latter has coverage probability

$$\begin{aligned}
\Pr\left(\theta \in \widehat{C}_1^0\right) &= \Pr\left(\widehat{\theta} + q(\alpha/2) \le \theta \le \widehat{\theta} + q(1-\alpha/2)\right) \\
&= \Pr\left(-q(1-\alpha/2) \le \widehat{\theta} - \theta \le -q(\alpha/2)\right) \\
&= G_n(-q(\alpha/2), F) - G_n(-q(1-\alpha/2), F)
\end{aligned}$$

which generally is not $1-\alpha$! There is one important exception. If $\widehat{\theta}-\theta$ has a symmetric distribution about 0, then $G_n(-u, F) = 1 - G_n(u, F)$, so

$$\begin{aligned}
\Pr\left(\theta \in \widehat{C}_1^0\right) &= G_n(-q(\alpha/2), F) - G_n(-q(1-\alpha/2), F) \\
&= (1 - G_n(q(\alpha/2), F)) - (1 - G_n(q(1-\alpha/2), F)) \\
&= \left(1 - \frac{\alpha}{2}\right) - \left(1 - \left(1 - \frac{\alpha}{2}\right)\right) \\
&= 1 - \alpha
\end{aligned}$$

and this idealized confidence interval is accurate. Therefore, $\widehat{C}_1^0$ and $\widehat{C}_1$ are designed for the case that $\widehat{\theta}$ has a symmetric distribution about θ.

When $\widehat{\theta}$ does not have a symmetric distribution, $\widehat{C}_1$ may perform quite poorly.

However, by the translation invariance argument presented above, it also follows that if there exists some monotonically increasing transformation $f(\cdot)$ such that $f(\widehat{\theta})$ is symmetrically distributed about $f(\theta)$, then the idealized percentile bootstrap method will be accurate.

Based on these arguments, many argue that the percentile interval should not be used unless the sampling distribution is close to unbiased and symmetric.

The problems with the percentile method can be circumvented, at least in principle, by an alternative method. Again, let $q(\alpha)$ and $q^*(\alpha)$ be the quantile functions of $\widehat{\theta} - \theta$ and $\widehat{\theta}^* - \widehat{\theta}$. Then

$$\begin{aligned}
1 - \alpha &= \Pr\left(q(\alpha/2) \le \widehat{\theta} - \theta \le q_n(1-\alpha/2)\right) \\
&= \Pr\left(\widehat{\theta} - q(1-\alpha/2) \le \theta \le \widehat{\theta} - q(\alpha/2)\right),
\end{aligned}$$

so an exact $1-\alpha$ confidence interval for θ is

$$\widehat{C}_2^0 = [\widehat{\theta} - q(1-\alpha/2), \quad \widehat{\theta} - q(\alpha/2)].$$

This motivates a bootstrap analog

$$\widehat{C}_2 = [\widehat{\theta} - q^*(1-\alpha/2), \quad \widehat{\theta} - q^*(\alpha/2)].$$

Notice that generally this is very different from the Efron interval $\widehat{C}_1$! They coincide in the special case that $G_n^*(u)$ is symmetric about $\widehat{\theta}$, but otherwise they differ.

Computationally, this interval can be estimated from a bootstrap simulation by sorting the bootstrap statistics $T^* = \widehat{\theta}^* - \widehat{\theta}$. These are sorted to yield the quantile estimates $\widehat{q}^*(.025)$ and $\widehat{q}^*(.975)$. The 95% confidence interval is then $[\widehat{\theta} - \widehat{q}^*(.975), \quad \widehat{\theta} - \widehat{q}^*(.025)]$.

This confidence interval is discussed in most theoretical treatments of the bootstrap, but is not widely used in practice.

13.6 Percentile-t Equal-Tailed Interval

Suppose we want to test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta < \theta_0$ at size α. We would set $T(\theta) = \left(\widehat{\theta} - \theta\right)/s(\widehat{\theta})$ and reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $T(\theta_0) < c$, where c would be selected so that

$$\Pr\left(T(\theta_0) < c\right) = \alpha.$$

Thus $c = q(\alpha)$. Since this is unknown, a bootstrap test replaces $q(\alpha)$ with the bootstrap estimate $q^*(\alpha)$, and the test rejects if $T(\theta_0) < q^*(\alpha)$.

Similarly, if the alternative is $\mathbb{H}_1 : \theta > \theta_0$, the bootstrap test rejects if $T(\theta_0) > q^*(1-\alpha)$.

Computationally, these critical values can be estimated from a bootstrap simulation by sorting the bootstrap t-statistics $T^* = \left(\widehat{\theta}^* - \widehat{\theta}\right)/s(\widehat{\theta}^*)$. Note, and this is important, that the bootstrap test statistic is centered at the estimate $\widehat{\theta}$, and the standard error $s(\widehat{\theta}^*)$ is calculated on the bootstrap sample. These t-statistics are sorted to find the estimated quantiles $\widehat{q}^*(\alpha)$ and/or $\widehat{q}^*(1-\alpha)$.

Let $T(\theta) = \left(\widehat{\theta} - \theta\right)/s(\widehat{\theta})$. Then taking the intersection of two one-sided intervals,

$$\begin{aligned}
1 - \alpha &= \Pr\left(q(\alpha/2) \le T(\theta_0) \le q(1-\alpha/2)\right) \\
&= \Pr\left(q(\alpha/2) \le \left(\widehat{\theta} - \theta_0\right)/s(\widehat{\theta}) \le q(1-\alpha/2)\right) \\
&= \Pr\left(\hat{\theta} - s(\widehat{\theta})q(1-\alpha/2) \le \theta_0 \le \hat{\theta} - s(\widehat{\theta})q(\alpha/2)\right).
\end{aligned}$$

An exact $(1-\alpha)\%$ confidence interval for θ is

$$\widehat{C}_3^0 = [\widehat{\theta} - s(\widehat{\theta})q(1-\alpha/2), \quad \widehat{\theta} - s(\widehat{\theta})q(\alpha/2)].$$

This motivates a bootstrap analog

$$\widehat{C}_3 = [\widehat{\theta} - s(\widehat{\theta})q^*(1-\alpha/2), \quad \widehat{\theta} - s(\widehat{\theta})q^*(\alpha/2)].$$

This is often called a **percentile-t confidence interval**. It is **equal-tailed** or **central** since the probability that θ is below the left endpoint approximately equals the probability that θ is above the right endpoint, each $\alpha/2$.

Computationally, this is based on the critical values from the one-sided hypothesis tests, discussed above.

13.7 Symmetric Percentile-t Intervals

Suppose we want to test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \ne \theta_0$ at size α. We would set $T(\theta) = \left(\widehat{\theta} - \theta\right)/s(\widehat{\theta})$ and reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $|T(\theta_0)| > c$, where c would be selected so that

$$\Pr\left(|T(\theta_0)| > c\right) = \alpha.$$

Note that

$$\begin{aligned}
\Pr\left(|T(\theta_0)| < c\right) &= \Pr\left(-c < T(\theta_0) < c\right) \\
&= G_n(c) - G_n(-c) \\
&\equiv \overline{G}_n(c),
\end{aligned}$$

which is a symmetric distribution function. The ideal critical value $c = q(1-\alpha)$ solves the equation

$$\overline{G}_n(q(1-\alpha)) = 1 - \alpha.$$

Equivalently, $q(1-\alpha)$ is the $1-\alpha$ quantile of the distribution of $|T(\theta_0)|$.

The bootstrap estimate is $q^*(1-\alpha)$, the $1-\alpha$ quantile of the distribution of $|T^*|$, or the number which solves the equation

$$\overline{G}_n^*(q^*(1-\alpha)) = G_n^*(q^*(1-\alpha)) - G_n^*(-q^*(1-\alpha)) = 1-\alpha.$$

Computationally, $q^*(1-\alpha)$ is estimated from a bootstrap simulation by sorting the bootstrap t-statistics $|T^*| = \left|\widehat{\theta}^* - \widehat{\theta}\right| / s(\widehat{\theta}^*)$, and taking the $1-\alpha$ quantile. The bootstrap test rejects if $|T(\theta_0)| > q^*(1-\alpha)$.

Let

$$\widehat{C}_4 = [\widehat{\theta} - s(\widehat{\theta})q^*(1-\alpha), \quad \widehat{\theta} + s(\widehat{\theta})q^*(1-\alpha)],$$

where $q_n^*(1-\alpha)$ is the bootstrap critical value for a two-sided hypothesis test. $\widehat{C}_4$ is called the symmetric percentile-t interval. It is designed to work well since

$$\begin{aligned}
\Pr\left(\theta \in \widehat{C}_4\right) &= \Pr\left(\widehat{\theta} - s(\widehat{\theta})q^*(1-\alpha) \le \theta \le \widehat{\theta} + s(\widehat{\theta})q^*(1-\alpha)\right) \\
&= \Pr\left(|T(\theta)| < q^*(1-\alpha)\right) \\
&\simeq \Pr\left(|T(\theta)| < q(1-\alpha)\right) \\
&= 1-\alpha.
\end{aligned}$$

If $\boldsymbol{\theta}$ is a vector, then to test $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ against $\mathbb{H}_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ at size α, we would use a Wald statistic

$$W(\boldsymbol{\theta}) = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$$

or a similar asymptotically chi-square statistic. The ideal test rejects if $W \ge q(1-\alpha)$, where $q(1-\alpha)$ is the $1-\alpha$ quantile of the distribution of W. The bootstrap test rejects if $W \ge q^*(1-\alpha)$, where $q^*(1-\alpha)$ is the $1-\alpha$ quantile of the distribution of

$$W^* = n\left(\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}}\right)' \widehat{\boldsymbol{V}}_{\theta}^{*-1} \left(\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}}\right).$$

Computationally, the critical value $q^*(1-\alpha)$ is found as the quantile from simulated values of W^*. Note in the simulation that the Wald statistic is a quadratic form in $\left(\widehat{\boldsymbol{\theta}}^* - \widehat{\boldsymbol{\theta}}\right)$, not $\left(\widehat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0\right)$. (The latter is a common mistake made by practitioners.)

13.8 Asymptotic Expansions

Let $T \in \mathbb{R}$ be a statistic such that

$$T \xrightarrow{d} \mathrm{N}(0, \sigma^2). \tag{13.3}$$

In some cases, such as when T is a t-ratio, then $\sigma^2 = 1$. In other cases σ^2 is unknown. Equivalently, writing $T \sim G_n(u, F)$ then for each u and F

$$\lim_{n\to\infty} G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right),$$

or

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + o(1). \tag{13.4}$$

While (13.4) says that G_n converges to $\Phi\left(\frac{u}{\sigma}\right)$ as $n \to \infty$, it says nothing, however, about the rate of convergence, or the size of the divergence for any particular sample size n. A better asymptotic approximation may be obtained through an **asymptotic expansion**.

Notationally, it is useful to recall the stochastic order notation of Section 6.13. Also, it is convenient to define even and odd functions. We say that a function $g(u)$ is **even** if $g(-u) = g(u)$, and a function $h(u)$ is **odd** if $h(-u) = -h(u)$. The derivative of an even function is odd, and vice-versa.

Theorem 13.8.1 *Under regularity conditions and (13.3),*

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + \frac{1}{n^{1/2}} g_1(u, F) + \frac{1}{n} g_2(u, F) + O(n^{-3/2})$$

uniformly over u, where g_1 is an even function of u, and g_2 is an odd function of u. Moreover, g_1 and g_2 are differentiable functions of u and continuous in F relative to the supremum norm on the space of distribution functions.

The expansion in Theorem 13.8.1 is often called an **Edgeworth expansion**.

We can interpret Theorem 13.8.1 as follows. First, $G_n(u, F)$ converges to the normal limit at rate $n^{1/2}$. To a second order of approximation,

$$G_n(u, F) \approx \Phi\left(\frac{u}{\sigma}\right) + n^{-1/2} g_1(u, F).$$

Since the derivative of g_1 is odd, the density function is skewed. To a third order of approximation,

$$G_n(u, F) \approx \Phi\left(\frac{u}{\sigma}\right) + n^{-1/2} g_1(u, F) + n^{-1} g_2(u, F)$$

which adds a symmetric non-normal component to the approximate density (for example, adding leptokurtosis).

As a side note, when $T = \sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma$, a standardized sample mean, then

$$g_1(u) = -\frac{1}{6}\kappa_3\left(u^2 - 1\right)\phi(u)$$
$$g_2(u) = -\left(\frac{1}{24}\kappa_4\left(u^3 - 3u\right) + \frac{1}{72}\kappa_3^2\left(u^5 - 10u^3 + 15u\right)\right)\phi(u)$$

where $\phi(u)$ is the standard normal pdf, and

$$\kappa_3 = \mathbb{E}\left(\left(X - \mu\right)^3\right)/\sigma^3$$
$$\kappa_4 = \mathbb{E}\left(\left(X - \mu\right)^4\right)/\sigma^4 - 3$$

the standardized skewness and excess kurtosis of the distribution of X. Note that when $\kappa_3 = 0$ and $\kappa_4 = 0$, then $g_1 = 0$ and $g_2 = 0$, so the second-order Edgeworth expansion corresponds to the normal distribution.

Francis Edgeworth

Francis Ysidro Edgeworth (1845-1926) of Ireland, founding editor of the *Economic Journal*, was a profound economic and statistical theorist, developing the theories of indifference curves and asymptotic expansions. He also could be viewed as the first econometrician due to his early use of mathematical statistics in the study of economic data.

13.9 One-Sided Tests

Using the expansion of Theorem 13.8.1, we can assess the accuracy of one-sided hypothesis tests and confidence regions based on an asymptotically normal t-ratio T. An asymptotic test is based on $\Phi(u)$.

To the second order, the exact distribution is

$$\Pr(T < u) = G_n(u, F) = \Phi(u) + \frac{1}{n^{1/2}} g_1(u, F) + O(n^{-1})$$

since $\sigma = 1$. The difference is

$$\begin{aligned} \Phi(u) - G_n(u, F) &= \frac{1}{n^{1/2}} g_1(u, F) + O(n^{-1}) \\ &= O(n^{-1/2}), \end{aligned}$$

so the order of the error is $O(n^{-1/2})$.

A bootstrap test is based on $G_n^*(u)$, which from Theorem 13.8.1 has the expansion

$$G_n^*(u) = G_n(u, \widehat{F}) = \Phi(u) + \frac{1}{n^{1/2}} g_1(u, \widehat{F}) + O(n^{-1}).$$

Because $\Phi(u)$ appears in both expansions, the difference between the bootstrap distribution and the true distribution is

$$G_n^*(u) - G_n(u, F) = \frac{1}{n^{1/2}} \left(g_1(u, \widehat{F}) - g_1(u, F) \right) + O(n^{-1}).$$

Since $\widehat{F}$ converges to F at rate $\sqrt{n}$, and g_1 is continuous with respect to F, the difference $\left(g_1(u, \widehat{F}) - g_1(u, F) \right)$ converges to 0 at rate $\sqrt{n}$. Heuristically,

$$\begin{aligned} g_1(u, \widehat{F}) - g_1(u, F) &\approx \frac{\partial}{\partial F} g_1(u, F) \left(\widehat{F} - F \right) \\ &= O(n^{-1/2}), \end{aligned}$$

The "derivative" $\frac{\partial}{\partial F} g_1(u, F)$ is only heuristic, as F is a function. We conclude that

$$G_n^*(u) - G_n(u, F) = O(n^{-1}),$$

or

$$\Pr(T^* \leq u) = \Pr(T \leq u) + O(n^{-1}),$$

which is an improved rate of convergence over the asymptotic test (which converged at rate $O(n^{-1/2})$). This rate can be used to show that one-tailed bootstrap inference based on the t-ratio achieves a so-called **asymptotic refinement** – the Type I error of the test converges at a faster rate than an analogous asymptotic test.

13.10 Symmetric Two-Sided Tests

If a random variable y has distribution function $H(u) = \Pr(y \leq u)$, then the random variable $|y|$ has distribution function

$$\overline{H}(u) = H(u) - H(-u)$$

since

$$\begin{aligned} \Pr(|y| \leq u) &= \Pr(-u \leq y \leq u) \\ &= \Pr(y \leq u) - \Pr(y \leq -u) \\ &= H(u) - H(-u). \end{aligned}$$

For example, if $Z \sim \mathrm{N}(0,1)$, then $|Z|$ has distribution function

$$\overline{\Phi}(u) = \Phi(u) - \Phi(-u) = 2\Phi(u) - 1.$$

Similarly, if T has exact distribution $G_n(u, F)$, then $|T|$ has the distribution function

$$\overline{G}_n(u, F) = G_n(u, F) - G_n(-u, F).$$

A two-sided hypothesis test rejects $\mathbb{H}_0$ for large values of $|T|$. Since $T \xrightarrow{d} Z$, then $|T| \xrightarrow{d} |Z| \sim \overline{\Phi}$. Thus asymptotic critical values are taken from the $\overline{\Phi}$ distribution, and exact critical values are taken from the $\overline{G}_n(u, F)$ distribution. From Theorem 13.8.1, we can calculate that

$$\begin{aligned}
\overline{G}_n(u, F) &= G_n(u, F) - G_n(-u, F) \\
&= \left(\Phi(u) + \frac{1}{n^{1/2}} g_1(u, F) + \frac{1}{n} g_2(u, F)\right) \\
&- \left(\Phi(-u) + \frac{1}{n^{1/2}} g_1(-u, F) + \frac{1}{n} g_2(-u, F)\right) + O(n^{-3/2}) \\
&= \overline{\Phi}(u) + \frac{2}{n} g_2(u, F) + O(n^{-3/2}), \qquad (13.5)
\end{aligned}$$

where the simplifications are because g_1 is even and g_2 is odd. Hence the difference between the asymptotic distribution and the exact distribution is

$$\overline{\Phi}(u) - \overline{G}_n(u, F_0) = \frac{2}{n} g_2(u, F_0) + O(n^{-3/2}) = O(n^{-1}).$$

The order of the error is $O(n^{-1})$.

Interestingly, the asymptotic two-sided test has a better coverage rate than the asymptotic one-sided test. This is because the first term in the asymptotic expansion, g_1, is an even function, meaning that the errors in the two directions exactly cancel out.

Applying (13.5) to the bootstrap distribution, we find

$$\overline{G}_n^*(u) = \overline{G}_n(u, \widehat{F}) = \overline{\Phi}(u) + \frac{2}{n} g_2(u, \widehat{F}) + O(n^{-3/2}).$$

Thus the difference between the bootstrap and exact distributions is

$$\begin{aligned}
\overline{G}_n^*(u) - \overline{G}_n(u, F) &= \frac{2}{n}\left(g_2(u, \widehat{F}) - g_2(u, F)\right) + O(n^{-3/2}) \\
&= O(n^{-3/2}),
\end{aligned}$$

the last equality because $\widehat{F}$ converges to F at rate $\sqrt{n}$, and g_2 is continuous in F. Another way of writing this is

$$\Pr\left(|T^*| < u\right) = \Pr\left(|T| < u\right) + O(n^{-3/2})$$

so the error from using the bootstrap distribution (relative to the true unknown distribution) is $O(n^{-3/2})$. This is in contrast to the use of the asymptotic distribution, whose error is $O(n^{-1})$. Thus a two-sided bootstrap test also achieves an asymptotic refinement, similar to a one-sided test.

A reader might get confused between the two simultaneous effects. Two-sided tests have better rates of convergence than the one-sided tests, and bootstrap tests have better rates of convergence than asymptotic tests.

The analysis shows that there may be a trade-off between one-sided and two-sided tests. Two-sided tests will have more accurate size (Reported Type I error), but one-sided tests might have more power against alternatives of interest. Confidence intervals based on the bootstrap can be asymmetric if based on one-sided tests (equal-tailed intervals) and can therefore be more informative and have smaller length than symmetric intervals. Therefore, the choice between symmetric and equal-tailed confidence intervals is unclear, and needs to be determined on a case-by-case basis.

13.11 Percentile Confidence Intervals

To evaluate the coverage rate of the percentile interval, set $T = \sqrt{n}\left(\widehat{\theta} - \theta\right)$. We know that $T \xrightarrow{d} \mathrm{N}(0, V)$, which is not pivotal, as it depends on the unknown V. Theorem 13.8.1 shows that a first-order approximation

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + O(n^{-1/2}),$$

where $\sigma = \sqrt{V}$, and for the bootstrap

$$G_n^*(u) = G_n(u, \widehat{F}) = \Phi\left(\frac{u}{\widehat{\sigma}}\right) + O(n^{-1/2}),$$

where $\widehat{\sigma} = V(\widehat{F})$ is the bootstrap estimate of σ. The difference is

$$\begin{aligned} G_n^*(u) - G_n(u, F) &= \Phi\left(\frac{u}{\widehat{\sigma}}\right) - \Phi\left(\frac{u}{\sigma}\right) + O(n^{-1/2}) \\ &= -\phi\left(\frac{u}{\sigma}\right)\frac{u}{\sigma}\left(\widehat{\sigma} - \sigma\right) + O(n^{-1/2}) \\ &= O(n^{-1/2}) \end{aligned}$$

Hence the order of the error is $O(n^{-1/2})$.

The good news is that the percentile-type methods (if appropriately used) can yield $\sqrt{n}$-convergent asymptotic inference. Yet these methods do not require the calculation of standard errors! This means that in contexts where standard errors are not available or are difficult to calculate, the percentile bootstrap methods provide an attractive inference method.

The bad news is that the rate of convergence is disappointing. It is no better than the rate obtained from an asymptotic one-sided confidence region. Therefore if standard errors are available, it is unclear if there are any benefits from using the percentile bootstrap over simple asymptotic methods.

Based on these arguments, the theoretical literature (e.g. Hall, 1992, Horowitz, 2001) tends to advocate the use of the percentile-t bootstrap methods rather than percentile methods.

13.12 Bootstrap Methods for Regression Models

The bootstrap methods we have discussed have set $G_n^*(u) = G_n(u, \widehat{F})$, where $\widehat{F}$ is the EDF. Any other consistent estimate of F may be used to define a feasible bootstrap estimator. The advantage of the EDF is that it is fully nonparametric, it imposes no conditions, and works in nearly any context. But since it is fully nonparametric, it may be inefficient in contexts where more is known about F. We discuss bootstrap methods appropriate for the linear regression model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0. \end{aligned}$$

The non-parametric bootstrap resamples the observations $(y_i^*, \boldsymbol{x}_i^*)$ from the EDF, which implies

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i^{*\prime}\widehat{\boldsymbol{\beta}} + e_i^* \\ \mathbb{E}\left(\boldsymbol{x}_i^* e_i^*\right) &= \mathbf{0} \end{aligned}$$

but generally

$$\mathbb{E}\left(e_i^* \mid \boldsymbol{x}_i^*\right) \neq 0.$$

The bootstrap distribution does not impose the regression assumption, and is thus an inefficient estimator of the true distribution (when in fact the regression assumption is true.)

One approach to this problem is to impose the very strong assumption that the error ε_i is independent of the regressor $\boldsymbol{x}_i$. The advantage is that in this case it is straightforward to construct bootstrap distributions. The disadvantage is that the bootstrap distribution may be a poor approximation when the error is not independent of the regressors.

To impose independence, it is sufficient to sample the $\boldsymbol{x}_i^*$ and e_i^* independently, and then create $y_i^* = \boldsymbol{x}_i^{*\prime}\widehat{\boldsymbol{\beta}} + e_i^*$. There are different ways to impose independence. A non-parametric method is to sample the bootstrap errors e_i^* randomly from the OLS residuals $\{\widehat{e}_1, ..., \widehat{e}_n\}$. A parametric method is to generate the bootstrap errors e_i^* from a parametric distribution, such as the normal $e_i^* \sim \mathrm{N}(0, \widehat{\sigma}^2)$.

For the regressors $\boldsymbol{x}_i^*$, a nonparametric method is to sample the $\boldsymbol{x}_i^*$ randomly from the EDF or sample values $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$. A parametric method is to sample $\boldsymbol{x}_i^*$ from an estimated parametric distribution. A third approach sets $\boldsymbol{x}_i^* = \boldsymbol{x}_i$. This is equivalent to treating the regressors as **fixed in repeated samples**. If this is done, then all inferential statements are made conditionally on the observed values of the regressors, which is a valid statistical approach. It does not really matter, however, whether or not the $\boldsymbol{x}_i$ are really "fixed" or random.

The methods discussed above are unattractive for most applications in econometrics because they impose the stringent assumption that $\boldsymbol{x}_i$ and e_i are independent. Typically what is desirable is to impose only the regression condition $\mathbb{E}(e_i \mid \boldsymbol{x}_i) = 0$. Unfortunately this is a harder problem.

One proposal which imposes the regression condition without independence is the **Wild Bootstrap**. The idea is to construct a conditional distribution for e_i^* so that

$$\begin{aligned}\mathbb{E}(e_i^* \mid \boldsymbol{x}_i) &= 0 \\ \mathbb{E}\left(e_i^{*2} \mid \boldsymbol{x}_i\right) &= \widehat{e}_i^2 \\ \mathbb{E}\left(e_i^{*3} \mid \boldsymbol{x}_i\right) &= \widehat{e}_i^3.\end{aligned}$$

A conditional distribution with these features will preserve the main important features of the data. This can be achieved using a two-point distribution of the form

$$\begin{aligned}\Pr\left(e_i^* = \left(\frac{1+\sqrt{5}}{2}\right)\widehat{e}_i\right) &= \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \Pr\left(e_i^* = \left(\frac{1-\sqrt{5}}{2}\right)\widehat{e}_i\right) &= \frac{\sqrt{5}+1}{2\sqrt{5}}\end{aligned}$$

For each $\boldsymbol{x}_i$, you sample e_i^* using this two-point distribution.

13.13 Bootstrap GMM Inference

Consider an unconditional moment model

$$\mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\beta})) = 0$$

and let $\widehat{\boldsymbol{\beta}}$ be the 2SLS or GMM estimator of $\boldsymbol{\beta}$. Using the EDF of $\boldsymbol{w}_i = (y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$, we can apply bootstrap methods to compute estimates of the bias and variance of $\widehat{\boldsymbol{\beta}}$ and construct confidence intervals for $\boldsymbol{\beta}$, identically as in the regression model. However, caution should be applied when interpreting such results.

A straightforward application of the nonparametric bootstrap works in the sense of consistently achieving the first-order asymptotic distribution. This has been shown by Hahn (1996). However, it fails to achieve an asymptotic refinement when the model is over-identified, jeopardizing the theoretical justification for percentile-t methods. Furthermore, the bootstrap applied J test will yield the wrong answer.

The problem is that in the sample, $\widehat{\boldsymbol{\beta}}$ is the "true" value and yet $\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}) \neq 0$. Thus according to random variables $(y_i^*, \boldsymbol{z}_i^*, \boldsymbol{x}_i^*)$ drawn from the EDF F_n,

$$\mathbb{E}\left(\boldsymbol{g}_i(\widehat{\boldsymbol{\beta}})\right) = \overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}) \neq \mathbf{0}.$$

This means that $(y_i^*, \boldsymbol{z}_i^*, \boldsymbol{x}_i^*)$ do not satisfy the same moment conditions as the population distribution.

A correction suggested by Hall and Horowitz (1996) can solve the problem. Given the bootstrap sample $(\boldsymbol{y}^*, \boldsymbol{Z}^*, \boldsymbol{X}^*)$, define the bootstrap GMM criterion

$$J^*(\boldsymbol{\beta}) = n \cdot \left(\overline{\boldsymbol{g}}_n^*(\boldsymbol{\beta}) - \overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})\right)' \widehat{\boldsymbol{W}}^* \left(\overline{\boldsymbol{g}}_n^*(\boldsymbol{\beta}) - \overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})\right)$$

where $\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})$ is from the in-sample data, not from the bootstrap data.

Let $\widehat{\boldsymbol{\beta}}^*$ minimize $J^*(\boldsymbol{\beta})$, and define all statistics and tests accordingly. In the linear model, this implies that the bootstrap estimator is

$$\widehat{\boldsymbol{\beta}}^* = \left(\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\boldsymbol{W}^*\boldsymbol{Z}^{*\prime}\boldsymbol{X}^*\right)^{-1}\left(\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\widehat{\boldsymbol{W}}^*\left(\boldsymbol{Z}^{*\prime}\boldsymbol{y}^* - \boldsymbol{Z}'\widehat{\boldsymbol{e}}\right)\right).$$

where $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are the in-sample residuals. The bootstrap J statistic is $J^*(\widehat{\boldsymbol{\beta}}^*)$.

Exercises

Exercise 13.1 Let $\widehat{F}(\boldsymbol{x})$ denote the EDF of a random sample. Show that

$$\sqrt{n}\left(\widehat{F}(\boldsymbol{x}) - F(\boldsymbol{x})\right) \xrightarrow{d} \mathrm{N}\left(0, F(\boldsymbol{x})\left(1 - F(\boldsymbol{x})\right)\right).$$

Exercise 13.2 Take a random sample $\{y_1, ..., y_n\}$ with $\mu = \mathbb{E}(y_i)$ and $\sigma^2 = \mathrm{var}(y_i)$ and set $\overline{y}_n = n^{-1}\sum_{i=1}^n y_i$. Find the population moments $\mathbb{E}(\overline{y}_n)$ and $\mathrm{var}(\overline{y}_n)$. Let $\{y_1^*, ..., y_n^*\}$ be a random sample from the empirical distribution function and set $\overline{y}_n^* = n^{-1}\sum_{i=1}^n y_i^*$. Find the bootstrap moments $\mathbb{E}(\overline{y}_n^*)$ and $\mathrm{var}(\overline{y}_n^*)$.

Exercise 13.3 Consider the following bootstrap procedure for a regression of y_i on $\boldsymbol{x}_i$. Let $\widehat{\boldsymbol{\beta}}$ denote the OLS estimator from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$, and $\widehat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ the OLS residuals.

(a) Draw a random vector $(\boldsymbol{x}^*, e^*)$ from the pair $\{(\boldsymbol{x}_i, \widehat{e}_i) : i = 1, ..., n\}$. That is, draw a random integer i' from $[1, 2, ..., n]$, and set $\boldsymbol{x}^* = \boldsymbol{x}_{i'}$ and $e^* = \widehat{e}_{i'}$. Set $y^* = \boldsymbol{x}^{*\prime}\widehat{\boldsymbol{\beta}} + e^*$. Draw (with replacement) n such vectors, creating a random bootstrap data set $(\boldsymbol{y}^*, \boldsymbol{X}^*)$.

(b) Regress $\boldsymbol{y}^*$ on $\boldsymbol{X}^*$, yielding OLS estimates $\widehat{\boldsymbol{\beta}}^*$ and any other statistic of interest.

Show that this bootstrap procedure is (numerically) identical to the non-parametric bootstrap.

Exercise 13.4 Consider the following bootstrap procedure. Using the non-parametric bootstrap, generate bootstrap samples, calculate the estimate $\widehat{\theta}^*$ on these samples and then calculate

$$T^* = (\widehat{\theta}^* - \widehat{\theta})/s(\widehat{\theta}),$$

where $s(\widehat{\theta})$ is the standard error in the original data. Let $q^*(.05)$ and $q^*(.95)$ denote the 5% and 95% quantiles of T^*, and define the bootstrap confidence interval

$$\widehat{C} = \left[\widehat{\theta} - s(\widehat{\theta})q^*(.95), \quad \widehat{\theta} - s(\widehat{\theta})q^*(.05)\right].$$

Show that $\widehat{C}$ exactly equals the Alternative percentile interval (not the percentile-t interval).

Exercise 13.5 You want to test $\mathbb{H}_0 : \theta = 0$ against $\mathbb{H}_1 : \theta > 0$. The test for $\mathbb{H}_0$ is to reject if $T_n = \widehat{\theta}/s(\widehat{\theta}) > c$ where c is picked so that Type I error is α. You do this as follows. Using the non-parametric bootstrap, you generate bootstrap samples, calculate the estimates $\widehat{\theta}^*$ on these samples and then calculate

$$T^* = \widehat{\theta}^*/s(\widehat{\theta}^*).$$

Let $q^*(.95)$ denote the 95% quantile of T^*. You replace c with $q^*(.95)$, and thus reject $\mathbb{H}_0$ if $T_n = \widehat{\theta}/s(\widehat{\theta}) > q^*(.95)$. What is wrong with this procedure?

Exercise 13.6 Suppose that in an application, $\widehat{\theta} = 1.2$ and $s(\widehat{\theta}) = .2$. Using the non-parametric bootstrap, 1000 samples are generated from the bootstrap distribution, and $\widehat{\theta}^*$ is calculated on each sample. The $\widehat{\theta}^*$ are sorted, and the 2.5% and 97.5% quantiles of the $\widehat{\theta}^*$ are .75 and 1.3, respectively.

(a) Report the 95% Efron Percentile interval for θ.

(b) Report the 95% Alternative Percentile interval for θ.

(c) With the given information, can you report the 95% Percentile-t interval for θ?

Exercise 13.7 Consider the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i|\boldsymbol{x}_i\right) &= 0 \end{aligned}$$

with y_i scalar and $\boldsymbol{x}_i$ a k vector. You have a random sample $(y_i, \boldsymbol{x}_i : i = 1, ..., n)$. You are interested in estimating the regression function $m(\boldsymbol{x}) = E\left(y_i|\boldsymbol{x}_i = \boldsymbol{x}\right)$ at a fixed vector x and constructing a 95% confidence interval.

(a) Write the standard estimator and asymptotic confidence interval for $m(\boldsymbol{x})$.

(b) Describe the percentile bootstrap confidence interval for $m(\boldsymbol{x})$.

(c) Describe the percentile-t bootstrap confidence interval for $m(\boldsymbol{x})$.

Exercise 13.8 The observed data is $\{y_i, x_i\} \in \mathbb{R} \times \mathbb{R}^k$, $k > 1$, $i = 1, ..., n$. Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(x_i e_i\right) &= 0 \\ \mu_3 &= \mathbb{E}\left(e_i^3\right) \end{aligned}$$

(a) Write down an estimator for μ_3.

(b) Explain how to use the Efron percentile method to construct a 90% confidence interval for μ_3 in this specific model.

Exercise 13.9 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(x_i e_i\right) &= 0 \\ \mathbb{E}\left(e_i^2\right) &= \sigma^2 \end{aligned}$$

Describe the bootstrap percentile confidence interval for σ^2.

Exercise 13.10 The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= 0 \end{aligned}$$

with x_{2i} scalar. Describe how to test $\mathbb{H}_0 : \beta_2 = 0$ against $\mathbb{H}_1 : \beta_2 \neq 0$ using the nonparametric bootstrap.

Exercise 13.11 The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= 0 \end{aligned}$$

with both $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{1i}$ $k \times 1$. Describe how to test $\mathbb{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ against $\mathbb{H}_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$ using the nonparametric bootstrap.

Exercise 13.12 Suppose a PhD student has a sample $(y_i, x_i, z_i : i = 1, ..., n)$ and estimates by OLS the equation

$$y_i = z_i\widehat{\alpha} + x_i'\widehat{\beta} + \widehat{e}_i$$

where α is the coefficient of interest and she is interested in testing $\mathbb{H}_0 : \alpha = 0$ against $\mathbb{H}_1 : \alpha \neq 0$. She obtains $\widehat{\alpha} = 2.0$ with standard error $s(\widehat{\alpha}) = 1.0$ so the value of the t-ratio for $\mathbb{H}_0$ is $T = \widehat{\alpha}/s(\widehat{\alpha}) = 2.0$. To assess significance, the student decides to use the bootstrap. She uses the following algorithm

1. Samples (y_i^*, x_i^*, z_i^*) randomly from the observations. (Random sampling with replacement). Creates a random sample with n observations.

2. On this pseudo-sample, estimates the equation

$$y_i^* = z_i^* \hat{\alpha}^* + x_i^{*\prime} \hat{\beta}^* + \hat{e}_i^*$$

by OLS and computes standard errors, including $s(\widehat{\alpha}^*)$. The t-ratio for $\mathbb{H}_0$, $T^* = \widehat{\alpha}^*/s(\widehat{\alpha}^*)$ is computed and stored.

3. This is repeated $B = 9999$ times.

4. The 95% empirical quantile $\widehat{q}_{.95}^*$ of the bootstrap absolute t-ratios $|T^*|$ is computed. It is $\widehat{q}_{.95}^* = 3.5$.

5. The student notes that while $|T| = 2 > 1.96$ (and thus an asymptotic 5% size test rejects $\mathbb{H}_0$), $|T| = 2 < \widehat{q}_{.95}^* = 3.5$ and thus the bootstrap test does not reject $\mathbb{H}_0$. As the bootstrap is more reliable, the student concludes that $\mathbb{H}_0$ cannot be rejected in favor of $\mathbb{H}_1$.

Question: Do you agree with the student's method and reasoning? Do you see an error in her method?

Exercise 13.13 Take the model

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= 0 \end{aligned}$$

The parameter of interest is $\theta = \beta_1\beta_2$. Show how to construct a confidence interval for θ using the following three methods.

1. Asymptotic Theory

2. Percentile Bootstrap

3. Equal-Tailed Percentile-t Bootstrap.

Your answer should be specific to this problem, not general.

Exercise 13.14 Let y_i be iid, $\mu = \mathbb{E}(y_i) > 0$, and $\theta = \mu^{-1}$. Let $\widehat{\mu} = \overline{Y}_n$ be the sample mean and $\widehat{\theta} = \widehat{\mu}^{-1}$.

(a) Is $\widehat{\theta}$ unbiased for θ?

(b) If $\widehat{\theta}$ is biased, can you determine the direction of the bias $\mathbb{E}\left(\widehat{\theta} - \theta\right)$ (up or down)?

(c) Could the nonparametric bootstrap be used to estimate the bias? If so, explain how.

Exercise 13.15 Take the model

$$\begin{aligned} y_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= 0 \\ \theta &= \frac{\beta_1}{\beta_2} \end{aligned}$$

Assume that the observations (y_i, x_{1i}, x_{2i}) are i.i.d. across $i = 1, ..., n$. Describe how you would construct the percentile-t bootstrap confidence interval for θ.

Exercise 13.16 The model is iid data, $i = 1, ..., n$,

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \end{aligned}$$

Does the presence of conditional heteroskedasticity invalidate the application of the non-parametric bootstrap? Explain.

Exercise 13.17 The RESET specification test for nonlinearity in a random sample is the following.

The null hypothesis is a linear regression

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \end{aligned}$$

The parameter $\boldsymbol{\beta}$ is estimated by OLS yielding predicted values $\widehat{y}_i$. Then a second-stage least-squares regression is estimated including both $\boldsymbol{x}_i$ and $\widehat{y}_i$

$$y_i = \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}} + \left(\widehat{y}_i\right)^2 \widetilde{\gamma} + \widetilde{e}_i$$

The RESET test statistic R is the squared t-ratio on $\widetilde{\gamma}$.

A colleague suggests obtaining the critical value for the test using the bootstrap. He proposes the following bootstrap implementation.

- Draw n observations $(y_i^*, \boldsymbol{x}_i^*)$ randomly from the observed sample pairs $(y_i, \boldsymbol{x}_i)$ to create a bootstrap sample.
- Compute the statistic R^* on this bootstrap sample as described above.
- Repeat this 999 times. Sort the bootstrap statistics R^*, take number 950 (the 95% percentile) and use this as the critical value.
- Reject the null hypothesis if R exceeds this critical value, otherwise do not reject.

Is this procedure a correct implementation of the bootstrap in this context? If not, propose a modified bootstrap.

Exercise 13.18 The model is

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &\neq 0, \end{aligned}$$

so the regressor $\boldsymbol{x}_i$ is endogenous. We know that in this case, the OLS estimator is biased for the parameter $\boldsymbol{\beta}$. We also know that the non-parametric bootstrap is (generally) a good method to estimate bias, and thereby make bias-adjusted. Explain whether or not the non-parametric bootstrap can be used to estimate the bias of OLS in the above context.

Exercise 13.19 The datafile `hprice1.txt` contains data on house prices (sales), with variables listed in the file `hprice1.pdf`. Estimate a linear regression of price on the number of bedrooms, lot size, size of house, and the colonial dummy. Calculate 95% confidence intervals for the regression coefficients using both the asymptotic normal approximation and the percentile-t bootstrap.

Chapter 14

Univariate Time Series

A time series y_t is a process observed in sequence over time, $t = 1, ..., T$. To indicate the dependence on time, we adopt new notation, and use the subscript t to denote the individual observation, and T to denote the number of observations.

Because of the sequential nature of time series, we expect that y_t and y_{t-1} are *not* independent, so classical assumptions are not valid.

We can separate time series into two categories: univariate ($y_t \in \mathbb{R}$ is scalar); and multivariate ($y_t \in \mathbb{R}^m$ is vector-valued). The primary model for univariate time series is autoregressions (ARs). The primary model for multivariate time series is vector autoregressions (VARs).

14.1 Stationarity and Ergodicity

Definition 14.1.1 *$\{y_t\}$ is **covariance (weakly) stationary** if*

$$\mathbb{E}(y_t) = \mu$$

is independent of t, and

$$\operatorname{cov}(y_t, y_{t-k}) = \gamma(k)$$

*is independent of t for all k.$\gamma(k)$ is called the **autocovariance function**.*

$$\rho(k) = \gamma(k)/\gamma(0) = \operatorname{corr}(y_t, y_{t-k})$$

*is the **autocorrelation function**.*

Definition 14.1.2 *$\{y_t\}$ is **strictly stationary** if the joint distribution of $(y_t, ..., y_{t-k})$ is independent of t for all k.*

Definition 14.1.3 *A stationary time series is **ergodic** if $\gamma(k) \to 0$ as $k \to \infty$.*

The following two theorems are essential to the analysis of stationary time series. The proofs are rather difficult, however.

Theorem 14.1.1 *If y_t is strictly stationary and ergodic and $x_t = f(y_t, y_{t-1}, ...)$ is a random variable, then x_t is strictly stationary and ergodic.*

Theorem 14.1.2 *(Ergodic Theorem). If y_t is strictly stationary and ergodic and $\mathbb{E}|y_t| < \infty$, then as $T \to \infty$,*

$$\frac{1}{T}\sum_{t=1}^{T} y_t \xrightarrow{p} \mathbb{E}(y_t).$$

This allows us to consistently estimate parameters using time-series moments:
The sample mean:

$$\widehat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

The sample autocovariance

$$\widehat{\gamma}(k) = \frac{1}{T}\sum_{t=1}^{T} (y_t - \widehat{\mu})(y_{t-k} - \widehat{\mu}).$$

The sample autocorrelation

$$\widehat{\rho}((k) = \frac{\widehat{\gamma}((k)}{\widehat{\gamma}((0)}.$$

Theorem 14.1.3 *If y_t is strictly stationary and ergodic and $\mathbb{E}\left(y_t^2\right) < \infty$, then as $T \to \infty$,*

1. $\widehat{\mu} \xrightarrow{p} \mathbb{E}(y_t)$;
2. $\widehat{\gamma}(k) \xrightarrow{p} \gamma(k)$;
3. $\widehat{\rho}(k) \xrightarrow{p} \rho(k)$.

Proof of Theorem 14.1.3. Part (1) is a direct consequence of the Ergodic theorem. For Part (2), note that

$$\begin{aligned}\widehat{\gamma}(k) &= \frac{1}{T}\sum_{t=1}^{T} (y_t - \widehat{\mu})(y_{t-k} - \widehat{\mu}) \\ &= \frac{1}{T}\sum_{t=1}^{T} y_t y_{t-k} - \frac{1}{T}\sum_{t=1}^{T} y_t\widehat{\mu} - \frac{1}{T}\sum_{t=1}^{T} y_{t-k}\widehat{\mu} + \widehat{\mu}^2.\end{aligned}$$

By Theorem 14.1.1 above, the sequence $y_t y_{t-k}$ is strictly stationary and ergodic, and it has a finite mean by the assumption that $\mathbb{E}\left(y_t^2\right) < \infty$. Thus an application of the Ergodic Theorem yields

$$\frac{1}{T}\sum_{t=1}^{T} y_t y_{t-k} \xrightarrow{p} \mathbb{E}(y_t y_{t-k}).$$

Thus

$$\widehat{\gamma}(k) \xrightarrow{p} \mathbb{E}(y_t y_{t-k}) - \mu^2 - \mu^2 + \mu^2 = \mathbb{E}(y_t y_{t-k}) - \mu^2 = \gamma(k).$$

Part (3) follows by the continuous mapping theorem: $\widehat{\rho}(k) = \widehat{\gamma}(k)/\widehat{\gamma}(0) \xrightarrow{p} \gamma(k)/\gamma(0) = \rho(k)$.

14.2 Autoregressions

In time-series, the series $\{..., y_1, y_2, ..., y_T, ...\}$ are jointly random. We consider the conditional expectation

$$\mathbb{E}\left(y_t \mid \mathcal{F}_{t-1}\right)$$

where $\mathcal{F}_{t-1} = \{y_{t-1}, y_{t-2}, ...\}$ is the past history of the series.

An autoregressive (AR) model specifies that only a finite number of past lags matter:

$$\mathbb{E}\left(y_t \mid \mathcal{F}_{t-1}\right) = \mathbb{E}\left(y_t \mid y_{t-1}, ..., y_{t-k}\right).$$

A linear AR model (the most common type used in practice) specifies linearity:

$$\mathbb{E}\left(y_t \mid \mathcal{F}_{t-1}\right) = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-1} + \cdots + \alpha_k y_{t-k}.$$

Letting

$$e_t = y_t - \mathbb{E}\left(y_t \mid \mathcal{F}_{t-1}\right),$$

then we have the autoregressive model

$$\begin{aligned} y_t &= \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-1} + \cdots + \alpha_k y_{t-k} + e_t \\ \mathbb{E}\left(e_t \mid \mathcal{F}_{t-1}\right) &= 0. \end{aligned}$$

The last property defines a special time-series process.

Definition 14.2.1 *e_t is a **martingale difference sequence** (MDS) if* $\mathbb{E}\left(e_t \mid \mathcal{F}_{t-1}\right) = 0$.

Regression errors are naturally a MDS. Some time-series processes may be a MDS as a consequence of optimizing behavior. For example, some versions of the life-cycle hypothesis imply that either changes in consumption, or consumption growth rates, should be a MDS. Most asset pricing models imply that asset returns should be the sum of a constant plus a MDS.

The MDS property for the regression error plays the same role in a time-series regression as does the conditional mean-zero property for the regression error in a cross-section regression. In fact, it is even more important in the time-series context, as it is difficult to derive distribution theories without this property.

A useful property of a MDS is that e_t is uncorrelated with any function of the lagged information $\mathcal{F}_{t-1}$. Thus for $k > 0$, $\mathbb{E}\left(y_{t-k} e_t\right) = 0$.

14.3 Stationarity of AR(1) Process

A mean-zero AR(1) is

$$y_t = \alpha y_{t-1} + e_t.$$

Assume that e_t is iid, $\mathbb{E}(e_t) = 0$ and $\mathbb{E}\left(e_t^2\right) = \sigma^2 < \infty$.

By back-substitution, we find

$$\begin{aligned} y_t &= e_t + \alpha e_{t-1} + \alpha^2 e_{t-2} + ... \\ &= \sum_{k=0}^{\infty} \alpha^k e_{t-k}. \end{aligned}$$

Loosely speaking, this series converges if the sequence $\alpha^k e_{t-k}$ gets small as $k \to \infty$. This occurs when $|\alpha| < 1$.

Theorem 14.3.1 *If and only if $|\alpha| < 1$ then y_t is strictly stationary and ergodic.*

We can compute the moments of y_t using the infinite sum:

$$\mathbb{E}(y_t) = \sum_{k=0}^{\infty} \alpha^k \mathbb{E}(e_{t-k}) = 0$$

$$\operatorname{var}(y_t) = \sum_{k=0}^{\infty} \alpha^{2k} \operatorname{var}(e_{t-k}) = \frac{\sigma^2}{1-\alpha^2}.$$

If the equation for y_t has an intercept, the above results are unchanged, except that the mean of y_t can be computed from the relationship

$$\mathbb{E}(y_t) = \alpha_0 + \alpha_1 \mathbb{E}(y_{t-1}),$$

and solving for $\mathbb{E}(y_t) = \mathbb{E}(y_{t-1})$ we find $\mathbb{E}(y_t) = \alpha_0/(1-\alpha_1)$.

14.4 Lag Operator

An algebraic construct which is useful for the analysis of autoregressive models is the lag operator.

Definition 14.4.1 *The **lag operator** L satisfies $\mathrm{L}y_t = y_{t-1}$.*

Defining $\mathrm{L}^2 = \mathrm{LL}$, we see that $\mathrm{L}^2 y_t = \mathrm{L}y_{t-1} = y_{t-2}$. In general, $\mathrm{L}^k y_t = y_{t-k}$.

The AR(1) model can be written in the format

$$y_t - \alpha y_{t-1} = e_t$$

or

$$(1 - \alpha \mathrm{L}) y_t = e_t.$$

The operator $\alpha(\mathrm{L}) = (1 - \alpha \mathrm{L})$ is a polynomial in the operator L. We say that the *root* of the polynomial is $1/\alpha$, since $\rho(z) = 0$ when $z = 1/\alpha$. We call $\alpha(\mathrm{L})$ the autoregressive polynomial of y_t.

From Theorem 14.3.1, an AR(1) is stationary iff $|\alpha| < 1$. Note that an equivalent way to say this is that an AR(1) is stationary iff the root of the autoregressive polynomial is larger than one (in absolute value).

14.5 Stationarity of AR(k)

The AR(k) model is

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \cdots + \alpha_k y_{t-k} + e_t.$$

Using the lag operator,

$$y_t - \alpha_1 \mathrm{L} y_t - \alpha_2 \mathrm{L}^2 y_t - \cdots - \alpha_k \mathrm{L}^k y_t = e_t,$$

or

$$\alpha(\mathrm{L}) y_t = e_t$$

where

$$\rho(\mathrm{L}) = 1 - \alpha_1 \mathrm{L} - \alpha_2 \mathrm{L}^2 - \cdots - \alpha_k \mathrm{L}^k.$$

We call $\alpha(\mathrm{L})$ the autoregressive polynomial of y_t.

The *Fundamental Theorem of Algebra* says that any polynomial can be factored as

$$\alpha(z) = \left(1 - \lambda_1^{-1} z\right)\left(1 - \lambda_2^{-1} z\right) \cdots \left(1 - \lambda_k^{-1} z\right)$$

where the $\lambda_1, ..., \lambda_k$ are the complex *roots* of $\alpha(z)$, which satisfy $\alpha(\lambda_j) = 0$.

We know that an AR(1) is stationary iff the absolute value of the root of its autoregressive polynomial is larger than one. For an AR(k), the requirement is that all roots are larger than one. Let $|\lambda|$ denote the modulus of a complex number λ.

Theorem 14.5.1 *The AR(k) is strictly stationary and ergodic if and only if $|\lambda_j| > 1$ for all j.*

One way of stating this is that "All roots lie outside the unit circle."

If one of the roots equals 1, we say that $\alpha(\mathrm{L})$, and hence y_t, "has a unit root". This is a special case of non-stationarity, and is of great interest in applied time series.

14.6 Estimation

Let

$$\boldsymbol{x}_t = \begin{pmatrix} 1 & y_{t-1} & y_{t-2} & \cdots & y_{t-k} \end{pmatrix}'$$
$$\boldsymbol{\beta} = \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \cdots & \alpha_k \end{pmatrix}'.$$

Then the model can be written as

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + e_t.$$

The OLS estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}.$$

To study $\widehat{\boldsymbol{\beta}}$, it is helpful to define the process $u_t = \boldsymbol{x}_t e_t$. Note that u_t is a MDS, since

$$\mathbb{E}\left(u_t \mid \mathcal{F}_{t-1}\right) = \mathbb{E}\left(\boldsymbol{x}_t e_t \mid \mathcal{F}_{t-1}\right) = \boldsymbol{x}_t \mathbb{E}\left(e_t \mid \mathcal{F}_{t-1}\right) = 0.$$

By Theorem 14.1.1, it is also strictly stationary and ergodic. Thus

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t e_t = \frac{1}{T}\sum_{t=1}^{T} u_t \xrightarrow{p} \mathbb{E}\left(u_t\right) = 0. \tag{14.1}$$

The vector $\boldsymbol{x}_t$ is strictly stationary and ergodic, and by Theorem 14.1.1, so is $\boldsymbol{x}_t\boldsymbol{x}_t'$. Thus by the Ergodic Theorem,

$$\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t' \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_t\boldsymbol{x}_t'\right) = \boldsymbol{Q}.$$

Combined with (14.1) and the continuous mapping theorem, we see that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t e_t\right) \xrightarrow{p} \boldsymbol{Q}^{-1}\boldsymbol{0} = \boldsymbol{0}.$$

We have shown the following:

Theorem 14.6.1 *If the AR(k) process y_t is strictly stationary and ergodic and $\mathbb{E}\left(y_t^2\right) < \infty$, then $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ as $T \to \infty$.*

14.7 Asymptotic Distribution

Theorem 14.7.1 *MDS CLT. If $\boldsymbol{u}_t$ is a strictly stationary and ergodic MDS and $\mathbb{E}\left(\boldsymbol{u}_t\boldsymbol{u}_t'\right) = \boldsymbol{\Omega} < \infty$, then as $T \to \infty$,*

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{u}_t \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right).$$

Since $\boldsymbol{x}_t e_t$ is a MDS, we can apply Theorem 14.7.1 to see that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\boldsymbol{x}_t e_t \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right),$$

where

$$\boldsymbol{\Omega} = \mathbb{E}(\boldsymbol{x}_t\boldsymbol{x}_t' e_t^2).$$

Theorem 14.7.2 *If the AR(k) process y_t is strictly stationary and ergodic and $\mathbb{E}\left(y_t^4\right) < \infty$, then as $T \to \infty$,*

$$\sqrt{T}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{Q}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}^{-1}\right).$$

This is identical in form to the asymptotic distribution of OLS in cross-section regression. The implication is that asymptotic inference is the same. In particular, the asymptotic covariance matrix is estimated just as in the cross-section case.

14.8 Bootstrap for Autoregressions

In the non-parametric bootstrap, we constructed the bootstrap sample by randomly resampling from the data values $\{y_t, \boldsymbol{x}_t\}$. This creates an iid bootstrap sample. Clearly, this cannot work in a time-series application, as this imposes inappropriate independence.

Briefly, there are two popular methods to implement bootstrap resampling for time-series data.

Method 1: Model-Based (Parametric) Bootstrap.

1. Estimate $\widehat{\boldsymbol{\beta}}$ and residuals $\widehat{e}_t$.
2. Fix an initial condition $(y_{-k+1}, y_{-k+2}, ..., y_0)$.
3. Simulate iid draws e_i^* from the empirical distribution of the residuals $\{\widehat{e}_1, ..., \widehat{e}_T\}$.
4. Create the bootstrap series y_t^* by the recursive formula

$$y_t^* = \widehat{\alpha}_0 + \widehat{\alpha}_1 y_{t-1}^* + \widehat{\alpha}_2 y_{t-2}^* + \cdots + \widehat{\alpha}_k y_{t-k}^* + e_t^*.$$

This construction imposes homoskedasticity on the errors e_i^*, which may be different than the properties of the actual e_i. It also presumes that the AR(k) structure is the truth.

Method 2: Block Resampling

1. Divide the sample into T/m blocks of length m.
2. Resample complete blocks. For each simulated sample, draw T/m blocks.
3. Paste the blocks together to create the bootstrap time-series y_t^*.
4. This allows for arbitrary stationary serial correlation, heteroskedasticity, and for model-misspecification.
5. The results may be sensitive to the block length, and the way that the data are partitioned into blocks.
6. May not work well in small samples.

14.9 Trend Stationarity

$$y_t = \mu_0 + \mu_1 t + S_t \tag{14.2}$$
$$S_t = \rho_1 S_{t-1} + \rho_2 S_{t-2} + \cdots + \rho_k S_{t-k} + e_t, \tag{14.3}$$

or

$$y_t = \alpha_0 + \alpha_1 t + \rho_1 y_{t-1} + \rho_2 y_{t-1} + \cdots + \rho_k y_{t-k} + e_t. \tag{14.4}$$

There are two essentially equivalent ways to estimate the autoregressive parameters $(\alpha_1, ..., \alpha_k)$.

- You can estimate (14.4) by OLS.
- You can estimate (14.2)-(14.3) sequentially by OLS. That is, first estimate (14.2), get the residual $\hat{S}_t$, and then perform regression (14.3) replacing S_t with $\hat{S}_t$. This procedure is sometimes called *Detrending*.

The reason why these two procedures are (essentially) the same is the Frisch-Waugh-Lovell theorem.

Seasonal Effects

There are three popular methods to deal with seasonal data.

- Include dummy variables for each season. This presumes that "seasonality" does not change over the sample.
- Use "seasonally adjusted" data. The seasonal factor is typically estimated by a two-sided weighted average of the data for that season in neighboring years. Thus the seasonally adjusted data is a "filtered" series. This is a flexible approach which can extract a wide range of seasonal factors. The seasonal adjustment, however, also alters the time-series correlations of the data.
- First apply a seasonal differencing operator. If s is the number of seasons (typically $s = 4$ or $s = 12$),
$$\Delta_s y_t = y_t - y_{t-s},$$
or the season-to-season change. The series $\Delta_s y_t$ is clearly free of seasonality. But the long-run trend is also eliminated, and perhaps this was of relevance.

14.10 Testing for Omitted Serial Correlation

For simplicity, let the null hypothesis be an AR(1):

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t. \tag{14.5}$$

We are interested in the question if the error u_t is serially correlated. We model this as an AR(1):

$$u_t = \theta u_{t-1} + e_t \tag{14.6}$$

with e_t a MDS. The hypothesis of no omitted serial correlation is

$$\mathbb{H}_0 : \theta = 0$$
$$\mathbb{H}_1 : \theta \neq 0.$$

We want to test $\mathbb{H}_0$ against $\mathbb{H}_1$.

To combine (14.5) and (14.6), we take (14.5) and lag the equation once:

$$y_{t-1} = \alpha_0 + \alpha_1 y_{t-2} + u_{t-1}.$$

We then multiply this by θ and subtract from (14.5), to find

$$y_t - \theta y_{t-1} = \alpha_0 - \theta\alpha_0 + \alpha_1 y_{t-1} - \theta\alpha_1 y_{t-1} + u_t - \theta u_{t-1},$$

or

$$y_t = \alpha_0(1-\theta) + (\alpha_1 + \theta)\, y_{t-1} - \theta\alpha_1 y_{t-2} + e_t = AR(2).$$

Thus under $\mathbb{H}_0$, y_t is an AR(1), and under $\mathbb{H}_1$ it is an AR(2). $\mathbb{H}_0$ may be expressed as the restriction that the coefficient on y_{t-2} is zero.

An appropriate test of $\mathbb{H}_0$ against $\mathbb{H}_1$ is therefore a Wald test that the coefficient on y_{t-2} is zero. (A simple exclusion test).

In general, if the null hypothesis is that y_t is an AR(k), and the alternative is that the error is an AR(m), this is the same as saying that under the alternative y_t is an AR(k+m), and this is equivalent to the restriction that the coefficients on $y_{t-k-1}, ..., y_{t-k-m}$ are jointly zero. An appropriate test is the Wald test of this restriction.

14.11 Model Selection

What is the appropriate choice of k in practice? This is a problem of model selection.

A good choice is to minimize the AIC information criterion

$$AIC(k) = \log \widehat{\sigma}^2(k) + \frac{2k}{T},$$

where $\widehat{\sigma}^2(k)$ is the estimated residual variance from an AR(k)

One ambiguity in defining the AIC criterion is that the sample available for estimation changes as k changes. (If you increase k, you need more initial conditions.) This can induce strange behavior in the AIC. The appropriate remedy is to fix a upper value $\overline{k}$, and then reserve the first $\overline{k}$ as initial conditions, and then estimate the models AR(1), AR(2), ..., AR($\overline{k}$) on this (unified) sample.

14.12 Autoregressive Unit Roots

The AR(k) model is

$$\begin{aligned} \alpha(\mathrm{L})y_t &= \alpha_0 + e_t \\ \alpha(\mathrm{L}) &= 1 - \alpha_1 \mathrm{L} - \cdots - \alpha_k \mathrm{L}^k. \end{aligned}$$

As we discussed before, y_t has a unit root when $\alpha(1) = 0$, or

$$\alpha_1 + \alpha_2 + \cdots + \alpha_k = 1.$$

In this case, y_t is non-stationary. The ergodic theorem and MDS CLT do not apply, and test statistics are asymptotically non-normal.

A helpful way to write the equation is the so-called Dickey-Fuller reparameterization:

$$\Delta y_t = \rho_0 y_{t-1} + \rho_1 \Delta y_{t-1} + \cdots + \rho_{k-1} \Delta y_{t-(k-1)} + e_t. \tag{14.7}$$

These models are equivalent linear transformations of one another. The DF parameterization is convenient because the parameter ρ_0 summarizes the information about the unit root, since $\alpha(1) = -\rho_0$. To see this, observe that the lag polynomial for the y_t computed from (14.7) is

$$(1 - \mathrm{L}) - \rho_0 \mathrm{L} - \rho_1(\mathrm{L} - \mathrm{L}^2) - \cdots - \rho_{k-1}(\mathrm{L}^{k-1} - \mathrm{L}^k)$$

But this must equal $\rho(\mathrm{L})$, as the models are equivalent. Thus

$$\alpha(1) = (1 - 1) - \rho_0 - (1 - 1) - \cdots - (1 - 1) = -\rho_0.$$

Hence, the hypothesis of a unit root in y_t can be stated as

$$\mathbb{H}_0 : \rho_0 = 0.$$

Note that the model is stationary if $\rho_0 < 0$. So the natural alternative is

$$\mathbb{H}_1 : \rho_0 < 0.$$

Under $\mathbb{H}_0$, the model for y_t is

$$\Delta y_t = \mu + \rho_1 \Delta y_{t-1} + \cdots + \rho_{k-1} \Delta y_{t-(k-1)} + e_t,$$

which is an AR(k-1) in the first-difference Δy_t. Thus if y_t has a (single) unit root, then Δy_t is a stationary AR process. Because of this property, we say that if y_t is non-stationary but $\Delta^d y_t$ is stationary, then y_t is "integrated of order d", or $I(d)$. Thus a time series with unit root is $I(1)$.

Since α_0 is the parameter of a linear regression, the natural test statistic is the t-statistic for $\mathbb{H}_0$ from OLS estimation of (14.7). Indeed, this is the most popular unit root test, and is called the Augmented Dickey-Fuller (ADF) test for a unit root.

It would seem natural to assess the significance of the ADF statistic using the normal table. However, under $\mathbb{H}_0$, y_t is non-stationary, so conventional normal asymptotics are invalid. An alternative asymptotic framework has been developed to deal with non-stationary data. We do not have the time to develop this theory in detail, but simply assert the main results.

Theorem 14.12.1 ***Dickey-Fuller Theorem.***
If $\rho_0 = 0$ *then as* $T \to \infty$,

$$T\widehat{\rho}_0 \xrightarrow{d} (1 - \rho_1 - \rho_2 - \cdots - \rho_{k-1})\, DF_\alpha$$

$$ADF = \frac{\hat{\rho}_0}{s(\hat{\rho}_0)} \to DF_t.$$

The limit distributions DF_α and DF_t are non-normal. They are skewed to the left, and have negative means.

The first result states that $\widehat{\rho}_0$ converges to its true value (of zero) at rate T, rather than the conventional rate of $T^{1/2}$. This is called a "super-consistent" rate of convergence.

The second result states that the t-statistic for $\widehat{\rho}_0$ converges to a limit distribution which is non-normal, but does not depend on the parameters ρ. This distribution has been extensively tabulated, and may be used for testing the hypothesis $\mathbb{H}_0$. Note: The standard error $s(\hat{\rho}_0)$ is the conventional ("homoskedastic") standard error. But the theorem does not require an assumption of homoskedasticity. Thus the Dickey-Fuller test is robust to heteroskedasticity.

Since the alternative hypothesis is one-sided, the ADF test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ when $ADF < c$, where c is the critical value from the ADF table. If the test rejects $\mathbb{H}_0$, this means that the evidence points to y_t being stationary. If the test does not reject $\mathbb{H}_0$, a common conclusion is that the data suggests that y_t is non-stationary. This is not really a correct conclusion, however. All we can say is that there is insufficient evidence to conclude whether the data are stationary or not.

We have described the test for the setting of with an intercept. Another popular setting includes as well a linear time trend. This model is

$$\Delta y_t = \mu_1 + \mu_2 t + \rho_0 y_{t-1} + \rho_1 \Delta y_{t-1} + \cdots + \rho_{k-1} \Delta y_{t-(k-1)} + e_t. \tag{14.8}$$

This is natural when the alternative hypothesis is that the series is stationary about a linear time trend. If the series has a linear trend (e.g. GDP, Stock Prices), then the series itself is non-stationary, but it may be stationary around the linear time trend. In this context, it is a silly waste of time to fit an AR model to the level of the series without a time trend, as the AR model cannot conceivably describe this data. The natural solution is to include a time trend in the fitted OLS equation. When conducting the ADF test, this means that it is computed as the t-ratio for ρ_0 from OLS estimation of (14.8).

If a time trend is included, the test procedure is the same, but different critical values are required. The ADF test has a different distribution when the time trend has been included, and a different table should be consulted.

Most texts include as well the critical values for the extreme polar case where the intercept has been omitted from the model. These are included for completeness (from a pedagogical perspective) but have no relevance for empirical practice where intercepts are always included.

Chapter 15

Multivariate Time Series

A multivariate time series $\boldsymbol{y}_t$ is a vector process $m \times 1$. Let $\mathcal{F}_{t-1} = (\boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, ...)$ be all lagged information at time t. The typical goal is to find the conditional expectation $\mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})$. Note that since $\boldsymbol{y}_t$ is a vector, this conditional expectation is also a vector.

15.1 Vector Autoregressions (VARs)

A VAR model specifies that the conditional mean is a function of only a finite number of lags:

$$\mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \mathbb{E}\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}\right).$$

A linear VAR specifies that this conditional mean is linear in the arguments:

$$\mathbb{E}\left(\boldsymbol{y}_t \mid \boldsymbol{y}_{t-1}, ..., \boldsymbol{y}_{t-k}\right) = \boldsymbol{a}_0 + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \cdots \boldsymbol{A}_k\boldsymbol{y}_{t-k}.$$

Observe that $\boldsymbol{a}_0$ is $m \times 1$,and each of $\boldsymbol{A}_1$ through $\boldsymbol{A}_k$ are $m \times m$ matrices.

Defining the $m \times 1$ regression error

$$e_t = \boldsymbol{y}_t - \mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}),$$

we have the VAR model

$$\begin{aligned} \boldsymbol{y}_t &= \boldsymbol{a}_0 + \boldsymbol{A}_1\boldsymbol{y}_{t-1} + \boldsymbol{A}_2\boldsymbol{y}_{t-2} + \cdots \boldsymbol{A}_k\boldsymbol{y}_{t-k} + \boldsymbol{e}_t \\ \mathbb{E}(\boldsymbol{e}_t \mid \mathcal{F}_{t-1}) &= \boldsymbol{0}. \end{aligned}$$

Alternatively, defining the $mk + 1$ vector

$$\boldsymbol{x}_t = \begin{pmatrix} 1 \\ \boldsymbol{y}_{t-1} \\ \boldsymbol{y}_{t-2} \\ \vdots \\ \boldsymbol{y}_{t-k} \end{pmatrix}$$

and the $m \times (mk + 1)$ matrix

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_0 & \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_k \end{pmatrix},$$

then

$$\boldsymbol{y}_t = \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{e}_t.$$

The VAR model is a system of m equations. One way to write this is to let a_j' be the jth row of $\boldsymbol{A}$. Then the VAR system can be written as the equations

$$Y_{jt} = a_j'\boldsymbol{x}_t + e_{jt}.$$

Unrestricted VARs were introduced to econometrics by Sims (1980).

15.2 Estimation

Consider the moment conditions

$$\mathbb{E}\left(\boldsymbol{x}_t e_{jt}\right) = \mathbf{0},$$

$j = 1, ..., m$. These are implied by the VAR model, either as a regression, or as a linear projection.

The GMM estimator corresponding to these moment conditions is equation-by-equation OLS

$$\widehat{\boldsymbol{a}}_j = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_j.$$

An alternative way to compute this is as follows. Note that

$$\widehat{\boldsymbol{a}}_j' = \boldsymbol{y}_j'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

And if we stack these to create the estimate $\widehat{\boldsymbol{A}}$, we find

$$\begin{aligned}\widehat{\boldsymbol{A}} &= \begin{pmatrix} \boldsymbol{y}_1' \\ \boldsymbol{y}_2' \\ \vdots \\ \boldsymbol{y}_{m+1}' \end{pmatrix} \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1},\end{aligned}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{y}_1 & \boldsymbol{y}_2 & \cdots & \boldsymbol{y}_m \end{pmatrix}$$

the $T \times m$ matrix of the stacked $\boldsymbol{y}_t'$.

This (system) estimator is known as the SUR (Seemingly Unrelated Regressions) estimator, and was originally derived by Zellner (1962)

15.3 Restricted VARs

The unrestricted VAR is a system of m equations, each with the same set of regressors. A restricted VAR imposes restrictions on the system. For example, some regressors may be excluded from some of the equations. Restrictions may be imposed on individual equations, or across equations. The GMM framework gives a convenient method to impose such restrictions on estimation.

15.4 Single Equation from a VAR

Often, we are only interested in a single equation out of a VAR system. This takes the form

$$y_{jt} = \boldsymbol{a}_j'\boldsymbol{x}_t + e_t,$$

and $\boldsymbol{x}_t$ consists of lagged values of y_{jt} and the other $y_{lt}'s$. In this case, it is convenient to re-define the variables. Let $y_t = y_{jt}$, and $\boldsymbol{z}_t$ be the other variables. Let $e_t = e_{jt}$ and $\beta = a_j$. Then the single equation takes the form

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t, \tag{15.1}$$

and

$$\boldsymbol{x}_t = \left[\begin{pmatrix} 1 & \boldsymbol{y}_{t-1} & \cdots & \boldsymbol{y}_{t-k} & \boldsymbol{z}_{t-1}' & \cdots & \boldsymbol{z}_{t-k}' \end{pmatrix}'\right].$$

This is just a conventional regression with time series data.

15.5 Testing for Omitted Serial Correlation

Consider the problem of testing for omitted serial correlation in equation (15.1). Suppose that e_t is an AR(1). Then

$$\begin{aligned} y_t &= \boldsymbol{x}_t'\boldsymbol{\beta} + e_t \\ e_t &= \theta e_{t-1} + u_t \\ \mathbb{E}\left(u_t \mid \mathcal{F}_{t-1}\right) &= 0. \end{aligned} \tag{15.2}$$

Then the null and alternative are

$$\mathbb{H}_0 : \theta = 0 \qquad \mathbb{H}_1 : \theta \neq 0.$$

Take the equation $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$, and subtract off the equation once lagged multiplied by θ, to get

$$\begin{aligned} y_t - \theta y_{t-1} &= \left(\boldsymbol{x}_t'\boldsymbol{\beta} + e_t\right) - \theta\left(\boldsymbol{x}_{t-1}'\boldsymbol{\beta} + e_{t-1}\right) \\ &= \boldsymbol{x}_t'\boldsymbol{\beta} - \theta\boldsymbol{x}_{t-1}\boldsymbol{\beta} + e_t - \theta e_{t-1}, \end{aligned}$$

or

$$y_t = \theta y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{x}_{t-1}'\boldsymbol{\gamma} + u_t, \tag{15.3}$$

which is a valid regression model.

So testing $\mathbb{H}_0$ versus $\mathbb{H}_1$ is equivalent to testing for the significance of adding $(y_{t-1}, \boldsymbol{x}_{t-1})$ to the regression. This can be done by a Wald test. We see that an appropriate, general, and simple way to test for omitted serial correlation is to test the significance of extra lagged values of the dependent variable and regressors.

You may have heard of the Durbin-Watson test for omitted serial correlation, which once was very popular, and is still routinely reported by conventional regression packages. The DW test is appropriate only when regression $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$ is not dynamic (has no lagged values on the RHS), and e_t is iid $\mathrm{N}(0, \sigma^2)$. Otherwise it is invalid.

Another interesting fact is that (15.2) is a special case of (15.3), under the restriction $\gamma = -\boldsymbol{\beta}\theta$. This restriction, which is called a common factor restriction, may be tested if desired. If valid, the model (15.2) may be estimated by iterated GLS. (A simple version of this estimator is called Cochrane-Orcutt.) Since the common factor restriction appears arbitrary, and is typically rejected empirically, direct estimation of (15.2) is uncommon in recent applications.

15.6 Selection of Lag Length in an VAR

If you want a data-dependent rule to pick the lag length k in a VAR, you may either use a testing-based approach (using, for example, the Wald statistic), or an information criterion approach. The formula for the AIC and BIC are

$$\begin{aligned} AIC(k) &= \log\det\left(\widehat{\boldsymbol{\Omega}}(k)\right) + 2\frac{p}{T} \\ BIC(k) &= \log\det\left(\widehat{\boldsymbol{\Omega}}(k)\right) + \frac{p\log(T)}{T} \\ \widehat{\boldsymbol{\Omega}}(k) &= \frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{e}}_t(k)\widehat{\boldsymbol{e}}_t(k)' \\ p &= m(km+1) \end{aligned}$$

where p is the number of parameters in the model, and $\widehat{\boldsymbol{e}}_t(k)$ is the OLS residual vector from the model with k lags. The log determinant is the criterion from the multivariate normal likelihood.

15.7 Granger Causality

Partition the data vector into $(\boldsymbol{y}_t, \boldsymbol{z}_t)$. Define the two information sets

$$\mathcal{F}_{1t} = \left(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, ...\right)$$
$$\mathcal{F}_{2t} = \left(\boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{y}_{t-1}, \boldsymbol{z}_{t-1}, \boldsymbol{y}_{t-2}, \boldsymbol{z}_{t-2}, , ...\right)$$

The information set $\mathcal{F}_{1t}$ is generated only by the history of $\boldsymbol{y}_t$, and the information set $\mathcal{F}_{2t}$ is generated by both $\boldsymbol{y}_t$ and $\boldsymbol{z}_t$. The latter has more information.

We say that $\boldsymbol{z}_t$ does not **Granger-cause** $\boldsymbol{y}_t$ if

$$\mathbb{E}\left(\boldsymbol{y}_t \mid \mathcal{F}_{1,t-1}\right) = \mathbb{E}\left(\boldsymbol{y}_t \mid \mathcal{F}_{2,t-1}\right).$$

That is, conditional on information in lagged $\boldsymbol{y}_t$, lagged $\boldsymbol{z}_t$ does not help to forecast $\boldsymbol{y}_t$. If this condition does not hold, then we say that $\boldsymbol{z}_t$ Granger-causes $\boldsymbol{y}_t$.

The reason why we call this "Granger Causality" rather than "causality" is because this is not a physical or structure definition of causality. If $\boldsymbol{z}_t$ is some sort of forecast of the future, such as a futures price, then $\boldsymbol{z}_t$ may help to forecast $\boldsymbol{y}_t$ even though it does not "cause" $\boldsymbol{y}_t$. This definition of causality was developed by Granger (1969) and Sims (1972).

In a linear VAR, the equation for $\boldsymbol{y}_t$ is

$$\boldsymbol{y}_t = \alpha + \rho_1 \boldsymbol{y}_{t-1} + \cdots + \rho_k \boldsymbol{y}_{t-k} + \boldsymbol{z}_{t-1}'\boldsymbol{\gamma}_1 + \cdots + \boldsymbol{z}_{t-k}'\boldsymbol{\gamma}_k + e_t.$$

In this equation, $\boldsymbol{z}_t$ does not Granger-cause $\boldsymbol{y}_t$ if and only if

$$\mathbb{H}_0 : \boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2 = \cdots = \boldsymbol{\gamma}_k = 0.$$

This may be tested using an exclusion (Wald) test.

This idea can be applied to blocks of variables. That is, $\boldsymbol{y}_t$ and/or $\boldsymbol{z}_t$ can be vectors. The hypothesis can be tested by using the appropriate multivariate Wald test.

If it is found that $\boldsymbol{z}_t$ does not Granger-cause $\boldsymbol{y}_t$, then we deduce that our time-series model of $\mathbb{E}\left(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}\right)$ does not require the use of $\boldsymbol{z}_t$. Note, however, that $\boldsymbol{z}_t$ may still be useful to explain other features of $\boldsymbol{y}_t$, such as the conditional variance.

Clive W. J. Granger

Clive Granger (1934-2009) of England was one of the leading figures in time-series econometrics, and co-winner in 2003 of the Nobel Memorial Prize in Economic Sciences (along with Robert Engle). In addition to formalizing the definition of causality known as Granger causality, he invented the concept of cointegration, introduced spectral methods into econometrics, and formalized methods for the combination of forecasts.

15.8 Cointegration

The idea of cointegration is due to Granger (1981), and was articulated in detail by Engle and Granger (1987).

Definition 15.8.1 *The $m \times 1$ series $\boldsymbol{y}_t$ is **cointegrated** if $\boldsymbol{y}_t$ is $I(1)$ yet there exists $\boldsymbol{\beta}$, $m \times r$, of rank r, such that $\boldsymbol{z}_t = \boldsymbol{\beta}'\boldsymbol{y}_t$ is $I(0)$. The r vectors in $\boldsymbol{\beta}$ are called the **cointegrating vectors**.*

If the series $\boldsymbol{y}_t$ is not cointegrated, then $r = 0$. If $r = m$, then $\boldsymbol{y}_t$ is $I(0)$. For $0 < r < m$, $\boldsymbol{y}_t$ is $I(1)$ and cointegrated.

In some cases, it may be believed that $\boldsymbol{\beta}$ is known a priori. Often, $\boldsymbol{\beta} = (1 \quad -1)'$. For example, if $\boldsymbol{y}_t$ is a pair of interest rates, then $\boldsymbol{\beta} = (1 \quad -1)'$ specifies that the spread (the difference in returns) is stationary. If $\boldsymbol{y} = (\log(C) \quad \log(I))'$, then $\boldsymbol{\beta} = (1 \quad -1)'$ specifies that $\log(C/I)$ is stationary.

In other cases, $\boldsymbol{\beta}$ may not be known.

If $\boldsymbol{y}_t$ is cointegrated with a single cointegrating vector ($r = 1$), then it turns out that $\boldsymbol{\beta}$ can be consistently estimated by an OLS regression of one component of $\boldsymbol{y}_t$ on the others. Thus $\boldsymbol{y}_t = (Y_{1t}, Y_{2t})$ and $\boldsymbol{\beta} = (\beta_1 \ \beta_2)$ and normalize $\beta_1 = 1$. Then $\widehat{\beta}_2 = (\boldsymbol{y}_2'\boldsymbol{y}_2)^{-1}\boldsymbol{y}_2'\boldsymbol{y}_1 \xrightarrow{p} \beta_2$. Furthermore this estimator is super-consistent: $T(\widehat{\beta}_2 - \beta_2) = O_p(1)$, as first shown by Stock (1987). While OLS is not, in general, a good method to estimate $\boldsymbol{\beta}$, it is useful in the construction of alternative estimators and tests.

We are often interested in testing the hypothesis of no cointegration:

$$\begin{aligned}\mathbb{H}_0 &: r = 0 \\ \mathbb{H}_1 &: r > 0.\end{aligned}$$

Suppose that $\boldsymbol{\beta}$ is known, so $\boldsymbol{z}_t = \boldsymbol{\beta}'\boldsymbol{y}_t$ is known. Then under $\mathbb{H}_0$ $\boldsymbol{z}_t$ is $I(1)$, yet under $\mathbb{H}_1$ $\boldsymbol{z}_t$ is $I(0)$. Thus $\mathbb{H}_0$ can be tested using a univariate ADF test on $\boldsymbol{z}_t$.

When $\boldsymbol{\beta}$ is unknown, Engle and Granger (1987) suggested using an ADF test on the estimated residual $\hat{z}_t = \widehat{\boldsymbol{\beta}}'\boldsymbol{y}_t$, from OLS of y_{1t} on y_{2t}. Their justification was Stock's result that $\widehat{\boldsymbol{\beta}}$ is super-consistent under $\mathbb{H}_1$. Under $\mathbb{H}_0$, however, $\widehat{\boldsymbol{\beta}}$ is not consistent, so the ADF critical values are not appropriate. The asymptotic distribution was worked out by Phillips and Ouliaris (1990).

When the data have time trends, it may be necessary to include a time trend in the estimated cointegrating regression. Whether or not the time trend is included, the asymptotic distribution of the test is affected by the presence of the time trend. The asymptotic distribution was worked out in B. Hansen (1992).

15.9 Cointegrated VARs

We can write a VAR as

$$\begin{aligned}\boldsymbol{A}(\mathrm{L})\boldsymbol{y}_t &= \boldsymbol{e}_t \\ \boldsymbol{A}(\mathrm{L}) &= \boldsymbol{I} - \boldsymbol{A}_1\mathrm{L} - \boldsymbol{A}_2\mathrm{L}^2 - \cdots - \boldsymbol{A}_k\mathrm{L}^k\end{aligned}$$

or alternatively as

$$\Delta\boldsymbol{y}_t = \boldsymbol{\Pi}\boldsymbol{y}_{t-1} + \boldsymbol{D}(\mathrm{L})\Delta\boldsymbol{y}_{t-1} + \boldsymbol{e}_t$$

where

$$\begin{aligned}\boldsymbol{\Pi} &= -\boldsymbol{A}(1) \\ &= -\boldsymbol{I} + \boldsymbol{A}_1 + \boldsymbol{A}_2 + \cdots + \boldsymbol{A}_k.\end{aligned}$$

Theorem 15.9.1 ***Granger Representation Theorem***
$\boldsymbol{y}_t$ is cointegrated with $m \times r$ $\boldsymbol{\beta}$ if and only if $\operatorname{rank}(\boldsymbol{\Pi}) = r$ *and* $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ *where α is $m \times r$,* $\operatorname{rank}(\boldsymbol{\alpha}) = r$.

Thus cointegration imposes a restriction upon the parameters of a VAR. The restricted model can be written as

$$\Delta \boldsymbol{y}_t = \boldsymbol{\alpha}\boldsymbol{\beta}'\boldsymbol{y}_{t-1} + \boldsymbol{D}(\mathrm{L})\Delta \boldsymbol{y}_{t-1} + \boldsymbol{e}_t$$
$$\Delta \boldsymbol{y}_t = \boldsymbol{\alpha}\boldsymbol{z}_{t-1} + \boldsymbol{D}(\mathrm{L})\Delta \boldsymbol{y}_{t-1} + \boldsymbol{e}_t.$$

If $\boldsymbol{\beta}$ is known, this can be estimated by OLS of $\Delta \boldsymbol{y}_t$ on $\boldsymbol{z}_{t-1}$ and the lags of $\Delta \boldsymbol{y}_t$.

If $\boldsymbol{\beta}$ is unknown, then estimation is done by "reduced rank regression", which is least-squares subject to the stated restriction. Equivalently, this is the MLE of the restricted parameters under the assumption that $\boldsymbol{e}_t$ is iid $\mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega})$.

One difficulty is that $\boldsymbol{\beta}$ is not identified without normalization. When $r = 1$, we typically just normalize one element to equal unity. When $r > 1$, this does not work, and different authors have adopted different identification schemes.

In the context of a cointegrated VAR estimated by reduced rank regression, it is simple to test for cointegration by testing the rank of $\boldsymbol{\Pi}$. These tests are constructed as likelihood ratio (LR) tests. As they were discovered by Johansen (1988, 1991, 1995), they are typically called the "Johansen Max and Trace" tests. Their asymptotic distributions are non-standard, and are similar to the Dickey-Fuller distributions.

Chapter 16

Panel Data

A panel is a set of observations on individuals, collected over time. An observation is the pair $\{y_{it}, \boldsymbol{x}_{it}\}$, where the i subscript denotes the individual, and the t subscript denotes time. A panel may be **balanced**:

$$\{y_{it}, \boldsymbol{x}_{it}\} : t = 1, ..., T; \quad i = 1, ..., n,$$

or **unbalanced**:

$$\{y_{it}, \boldsymbol{x}_{it}\} : \text{For } i = 1, ..., n, \quad t = \underline{t}_i, ..., \overline{t}_i.$$

16.1 Individual-Effects Model

The standard panel data specification is that there is an individual-specific effect which enters linearly in the regression

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + e_{it}.$$

The typical maintained assumptions are that the individuals i are mutually independent, that u_i and e_{it} are independent, that e_{it} is iid across individuals and time, and that e_{it} is uncorrelated with $\boldsymbol{x}_{it}$.

OLS of y_{it} on $\boldsymbol{x}_{it}$ is called **pooled estimation**. It is consistent if

$$\mathbb{E}(\boldsymbol{x}_{it} u_i) = 0 \tag{16.1}$$

If this condition fails, then OLS is inconsistent. (16.1) fails if the individual-specific unobserved effect u_i is correlated with the observed explanatory variables $\boldsymbol{x}_{it}$. This is often believed to be plausible if u_i is an omitted variable.

If (16.1) is true, however, OLS can be improved upon via a GLS technique. In either event, OLS appears a poor estimation choice.

Condition (16.1) is called the **random effects hypothesis**. It is a strong assumption, and most applied researchers try to avoid its use.

16.2 Fixed Effects

This is the most common technique for estimation of non-dynamic linear panel regressions.

The motivation is to allow u_i to be arbitrary, and have arbitrary correlated with $\boldsymbol{x}_i$. The goal is to eliminate u_i from the estimator, and thus achieve invariance.

There are several derivations of the estimator.

First, let

$$d_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases},$$

and

$$\boldsymbol{d}_i = \begin{pmatrix} d_{i1} \\ \vdots \\ d_{in} \end{pmatrix},$$

an $n \times 1$ dummy vector with a "1" in the i^{th} place. Let

$$\boldsymbol{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}.$$

Then note that

$$u_i = \boldsymbol{d}_i' \boldsymbol{u},$$

and

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \boldsymbol{d}_i'\boldsymbol{u} + e_{it}. \tag{16.2}$$

Observe that

$$\mathbb{E}\left(e_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{d}_i\right) = 0,$$

so (16.2) is a valid regression, with $\boldsymbol{d}_i$ as a regressor along with $\boldsymbol{x}_i$.

OLS on (16.2) yields estimator $\left(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{u}}\right)$. Conventional inference applies.

Observe that

- This is generally consistent.
- If $\boldsymbol{x}_{it}$ contains an intercept, it will be collinear with $\boldsymbol{d}_i$, so the intercept is typically omitted from $\boldsymbol{x}_{it}$.
- Any regressor in $\boldsymbol{x}_{it}$ which is constant over time for all individuals (e.g., their gender) will be collinear with $\boldsymbol{d}_i$, so will have to be omitted.
- There are $n + k$ regression parameters, which is quite large as typically n is very large.

Computationally, you do not want to actually implement conventional OLS estimation, as the parameter space is too large. OLS estimation of $\boldsymbol{\beta}$ proceeds by the FWL theorem. Stacking the observations together:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{D}\boldsymbol{u} + e,$$

then by the FWL theorem,

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{D}}\right)\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{D}}\right)\boldsymbol{y}\right) \\ &= \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^{*}\right)^{-1}\left(\boldsymbol{X}^{*\prime}\boldsymbol{y}^{*}\right), \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{y}^* &= \boldsymbol{y} - \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{y} \\ \boldsymbol{X}^* &= \boldsymbol{X} - \boldsymbol{D}(\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{D}'\boldsymbol{X}. \end{aligned}$$

Since the regression of y_{it} on $\boldsymbol{d}_i$ is a regression onto individual-specific dummies, the predicted value from these regressions is the individual specific mean $\overline{y}_i$, and the residual is the demean value

$$y_{it}^* = y_{it} - \overline{y}_i.$$

The fixed effects estimator $\widehat{\boldsymbol{\beta}}$ is OLS of y_{it}^* on $\boldsymbol{x}_{it}^*$, the dependent variable and regressors in deviation-from-mean form.

Another derivation of the estimator is to take the equation

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + e_{it},$$

and then take individual-specific means by taking the average for the i^{th} individual:

$$\frac{1}{T_i}\sum_{t=\underline{t}_i}^{\overline{t}_i} y_{it} = \frac{1}{T_i}\sum_{t=\underline{t}_i}^{\overline{t}_i} \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + \frac{1}{T_i}\sum_{t=\underline{t}_i}^{\overline{t}_i} e_{it}$$

or

$$\overline{y}_i = \overline{\boldsymbol{x}}_i'\boldsymbol{\beta} + u_i + \overline{e}_i.$$

Subtracting, we find

$$y_{it}^* = \boldsymbol{x}_{it}^{*\prime}\boldsymbol{\beta} + e_{it}^*,$$

which is free of the individual-effect u_i.

16.3 Dynamic Panel Regression

A dynamic panel regression has a lagged dependent variable

$$y_{it} = \alpha y_{it-1} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + e_{it}. \tag{16.3}$$

This is a model suitable for studying dynamic behavior of individual agents.

Unfortunately, the fixed effects estimator is inconsistent, at least if T is held finite as $n \to \infty$. This is because the sample mean of y_{it-1} is correlated with that of e_{it}.

The standard approach to estimate a dynamic panel is to combine first-differencing with IV or GMM. Taking first-differences of (16.3) eliminates the individual-specific effect:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta} + \Delta e_{it}. \tag{16.4}$$

However, if e_{it} is iid, then it will be correlated with Δy_{it-1} :

$$\mathbb{E}\left(\Delta y_{it-1}\Delta e_{it}\right) = \mathbb{E}\left(\left(y_{it-1} - y_{it-2}\right)\left(e_{it} - e_{it-1}\right)\right) = -\mathbb{E}\left(y_{it-1}e_{it-1}\right) = -\sigma_e^2.$$

So OLS on (16.4) will be inconsistent.

But if there are valid instruments, then IV or GMM can be used to estimate the equation. Typically, we use lags of the dependent variable, two periods back, as y_{t-2} is uncorrelated with Δe_{it}. Thus values of y_{it-k}, $k \geq 2$, are valid instruments.

Hence a valid estimator of α and $\boldsymbol{\beta}$ is to estimate (16.4) by IV using y_{t-2} as an instrument for Δy_{t-1} (which is just identified). Alternatively, GMM using y_{t-2} and y_{t-3} as instruments (which is overidentified, but loses a time-series observation).

A more sophisticated GMM estimator recognizes that for time-periods later in the sample, there are more instruments available, so the instrument list should be different for each equation. This is conveniently organized by the GMM principle, as this enables the moments from the different time-periods to be stacked together to create a list of all the moment conditions. A simple application of GMM yields the parameter estimates and standard errors.

Exercises

Exercise 16.1 Consider the model

$$\begin{aligned} y_{it} &= \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + e_{it} \\ \mathbb{E}\left(\boldsymbol{z}_{it} e_{it}\right) &= 0 \end{aligned}$$

for $i = 1, ..., n$ and $t = 1, ..., T$. The individual effect u_i is treated as fixed. Assume $\boldsymbol{x}_{it}$ and $\boldsymbol{z}_{it}$ are $k \times 1$ vectors.

Write out an appropriate estimator for $\boldsymbol{\beta}$.

Chapter 17

NonParametric Regression

17.1 Introduction

When components of $\boldsymbol{x}$ are continuously distributed then the conditional expectation function

$$\mathbb{E}\left(y_i \mid \boldsymbol{x}_i = \boldsymbol{x}\right) = m(\boldsymbol{x})$$

can take any nonlinear shape. Unless an economic model restricts the form of $m(\boldsymbol{x})$ to a parametric function, the CEF is inherently **nonparametric**, meaning that the function $m(\boldsymbol{x})$ is an element of an infinite dimensional class. In this situation, how can we estimate $m(\boldsymbol{x})$? What is a suitable method, if we acknowledge that $m(\boldsymbol{x})$ is nonparametric?

There are two main classes of nonparametric regression estimators: kernel estimators, and series estimators. In this chapter we introduce kernel methods.

To get started, suppose that there is a single real-valued regressor x_i. We consider the case of vector-valued regressors later.

17.2 Binned Estimator

For clarity, fix the point x and consider estimation of the single point $m(x)$. This is the mean of y_i for random pairs (y_i, x_i) such that $x_i = x$. If the distribution of x_i were discrete then we could estimate $m(x)$ by taking the average of the sub-sample of observations y_i for which $x_i = x$. But when x_i is continuous then the probability is zero that x_i exactly equals any specific x. So there is no sub-sample of observations with $x_i = x$ and we cannot simply take the average of the corresponding y_i values. However, if the CEF $m(x)$ is continuous, then it should be possible to get a good approximation by taking the average of the observations for which x_i is *close* to x, perhaps for the observations for which $|x_i - x| \leq h$ for some small $h > 0$. We call h a **bandwidth**. This estimator can be written as

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} 1\left(|x_i - x| \leq h\right) y_i}{\sum_{i=1}^{n} 1\left(|x_i - x| \leq h\right)} \tag{17.1}$$

where $1(\cdot)$ is the indicator function. Alternatively, (17.1) can be written as

$$\widehat{m}(x) = \sum_{i=1}^{n} w_i(x) y_i \tag{17.2}$$

where

$$w_i(x) = \frac{1\left(|x_i - x| \leq h\right)}{\sum_{j=1}^{n} 1\left(|x_j - x| \leq h\right)}.$$

Notice that $\sum_{i=1}^{n} w_i(x) = 1$, so (17.2) is a weighted average of the y_i.

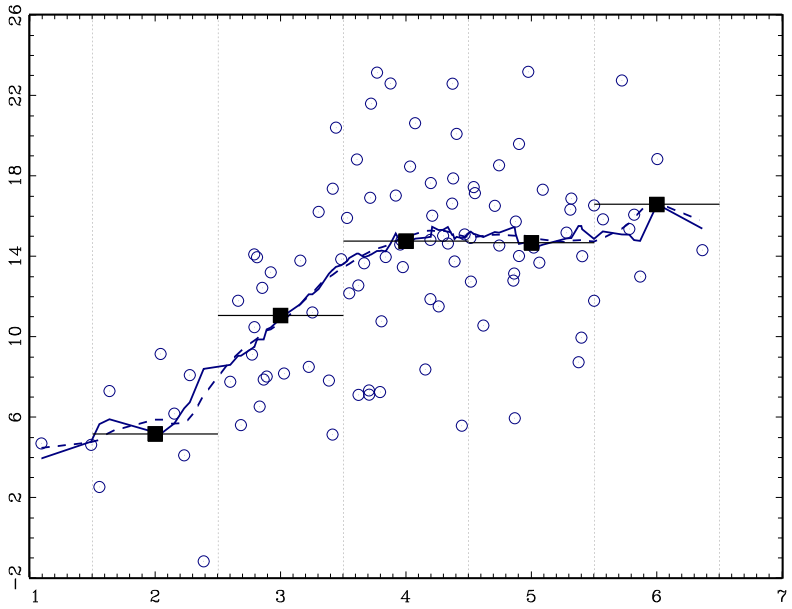


Figure 17.1: Scatter of (y_i, x_i) and Nadaraya-Watson regression

It is possible that for some values of x there are no values of x_i such that $|x_i - x| \leq h$, which implies that $\sum_{i=1}^{n} 1\left(|x_i - x| \leq h\right) = 0$. In this case the estimator (17.1) is undefined for those values of x.

To visualize, Figure 17.1 displays a scatter plot of 100 observations on a random pair (y_i, x_i) generated by simulation[1]. (The observations are displayed as the open circles.) The estimator (17.1) of the CEF $m(x)$ at $x = 2$ with $h = 1/2$ is the average of the y_i for the observations such that x_i falls in the interval $[1.5 \leq x_i \leq 2.5]$. (Our choice of $h = 1/2$ is somewhat arbitrary. Selection of h will be discussed later.) The estimate is $\widehat{m}(2) = 5.16$ and is shown on Figure 17.1 by the first solid square. We repeat the calculation (17.1) for $x = 3$, 4, 5, and 6, which is equivalent to partitioning the support of x_i into the regions $[1.5, 2.5]$, $[2.5, 3.5]$, $[3.5, 4.5]$, $[4.5, 5.5]$, and $[5.5, 6.5]$. These partitions are shown in Figure 17.1 by the verticle dotted lines, and the estimates (17.1) by the solid squares.

These estimates $\widehat{m}(x)$ can be viewed as estimates of the CEF $m(x)$. Sometimes called a binned estimator, this is a step-function approximation to $m(x)$ and is displayed in Figure 17.1 by the horizontal lines passing through the solid squares. This estimate roughly tracks the central tendency of the scatter of the observations (y_i, x_i). However, the huge jumps in the estimated step function at the edges of the partitions are disconcerting, counter-intuitive, and clearly an artifact of the discrete binning.

If we take another look at the estimation formula (17.1) there is no reason why we need to evaluate (17.1) only on a course grid. We can evaluate $\widehat{m}(x)$ for any set of values of x. In particular, we can evaluate (17.1) on a fine grid of values of x and thereby obtain a smoother estimate of the CEF. This estimator with $h = 1/2$ is displayed in Figure 17.1 with the solid line. This is a generalization of the binned estimator and by construction passes through the solid squares.

The bandwidth h determines the degree of smoothing. Larger values of h increase the width of the bins in Figure 17.1, thereby increasing the smoothness of the estimate $\widehat{m}(x)$ as a function of x. Smaller values of h decrease the width of the bins, resulting in less smooth conditional mean estimates.

[1] The distribution is $x_i \sim N(4, 1)$ and $y_i \mid x_i \sim N(m(x_i), 16)$ with $m(x) = 10\log(x)$.

17.3 Kernel Regression

One deficiency with the estimator (17.1) is that it is a step function in x, as it is discontinuous at each observation $x = x_i$. That is why its plot in Figure 17.1 is jagged. The source of the discontinuity is that the weights $w_i(x)$ are constructed from indicator functions, which are themselves discontinuous. If instead the weights are constructed from continuous functions then the CEF estimator will also be continuous in x.

To generalize (17.1) it is useful to write the weights $1(|x_i - x| \leq h)$ in terms of the uniform density function on $[-1, 1]$

$$k_0(u) = \frac{1}{2} 1(|u| \leq 1).$$

Then

$$1(|x_i - x| \leq h) = 1\left(\left|\frac{x_i - x}{h}\right| \leq 1\right) = 2k_0\left(\frac{x_i - x}{h}\right).$$

and (17.1) can be written as

$$\widehat{m}(x) = \frac{\sum_{i=1}^n k_0\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^n k_0\left(\frac{x_i - x}{h}\right)}. \tag{17.3}$$

The uniform density $k_0(u)$ is a special case of what is known as a **kernel function**.

Definition 17.3.1 *A second-order* ***kernel function*** $k(u)$ *satisfies* $0 \leq k(u) < \infty$, $k(u) = k(-u)$, $\int_{-\infty}^{\infty} k(u)du = 1$ *and* $\sigma_k^2 = \int_{-\infty}^{\infty} u^2 k(u)du < \infty$.

Essentially, a kernel function is a probability density function which is bounded and symmetric about zero. A generalization of (17.1) is obtained by replacing the uniform kernel with any other kernel function:

$$\widehat{m}(x) = \frac{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)}. \tag{17.4}$$

The estimator (17.4) also takes the form (17.2) with

$$w_i(x) = \frac{k\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^n k\left(\frac{x_j - x}{h}\right)}.$$

The estimator (17.4) is known as the **Nadaraya-Watson** estimator, the **kernel regression** estimator, or the **local constant** estimator.

The bandwidth h plays the same role in (17.4) as it does in (17.1). Namely, larger values of h will result in estimates $\widehat{m}(x)$ which are smoother in x, and smaller values of h will result in estimates which are more erratic. It might be helpful to consider the two extreme cases $h \to 0$ and $h \to \infty$. As $h \to 0$ we can see that $\widehat{m}(x_i) \to y_i$ (if the values of x_i are unique), so that $\widehat{m}(x)$ is simply the scatter of y_i on x_i. In contrast, as $h \to \infty$ then for all x, $\widehat{m}(x) \to \overline{y}$, the sample mean, so that the nonparametric CEF estimate is a constant function. For intermediate values of h, $\widehat{m}(x)$ will lie between these two extreme cases.

The uniform density is not a good kernel choice as it produces discontinuous CEF estimates. To obtain a continuous CEF estimate $\widehat{m}(x)$ it is necessary for the kernel $k(u)$ to be continuous. The two most commonly used choices are the **Epanechnikov kernel**

$$k_1(u) = \frac{3}{4}\left(1-u^2\right)1\left(|u| \leq 1\right)$$

and the **normal** or **Gaussian kernel**

$$k_\phi(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right).$$

For computation of the CEF estimate (17.4) the scale of the kernel is not important so long as the bandwidth is selected appropriately. That is, for any $b > 0$, $k_b(u) = b^{-1}k\left(\frac{u}{b}\right)$ is a valid kernel function with the identical shape as $k(u)$. Kernel regression with the kernel $k(u)$ and bandwidth h is identical to kernel regression with the kernel $k_b(u)$ and bandwidth h/b.

The estimate (17.4) using the Epanechnikov kernel and $h = 1/2$ is also displayed in Figure 17.1 with the dashed line. As you can see, this estimator appears to be much smoother than that using the uniform kernel.

Two important constants associated with a kernel function $k(u)$ are its variance σ_k^2 and roughness R_k, which are defined as

$$\sigma_k^2 = \int_{-\infty}^{\infty} u^2 k(u)du \tag{17.5}$$

$$R_k = \int_{-\infty}^{\infty} k(u)^2 du. \tag{17.6}$$

Some common kernels and their roughness and variance values are reported in Table 9.1.

Table 9.1: Common Second-Order Kernels

Kernel	**Equation**	R_k	σ_k^2
Uniform	$k_0(u) = \frac{1}{2}1\left(\lvert u\rvert \leq 1\right)$	$1/2$	$1/3$
Epanechnikov	$k_1(u) = \frac{3}{4}\left(1-u^2\right)1\left(\lvert u\rvert \leq 1\right)$	$3/5$	$1/5$
Biweight	$k_2(u) = \frac{15}{16}\left(1-u^2\right)^2 1\left(\lvert u\rvert \leq 1\right)$	$5/7$	$1/7$
Triweight	$k_3(u) = \frac{35}{32}\left(1-u^2\right)^3 1\left(\lvert u\rvert \leq 1\right)$	$350/429$	$1/9$
Gaussian	$k_\phi(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$	$1/\left(2\sqrt{\pi}\right)$	1

17.4 Local Linear Estimator

The Nadaraya-Watson (NW) estimator is often called a local constant estimator as it locally (about x) approximates the CEF $m(x)$ as a constant function. One way to see this is to observe that $\widehat{m}(x)$ solves the minimization problem

$$\widehat{m}(x) = \underset{\alpha}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)\left(y_i - \alpha\right)^2.$$

This is a weighted regression of y_i on an intercept only. Without the weights, this estimation problem reduces to the sample mean. The NW estimator generalizes this to a local mean.

This interpretation suggests that we can construct alternative nonparametric estimators of the CEF by alternative local approximations. Many such local approximations are possible. A popular choice is the **Local Linear** (LL) approximation. Instead of approximating $m(x)$ locally

as a constant, LL approximates the CEF locally by a linear function, and estimates this local approximation by locally weighted least squares.

Specifically, for each x we solve the following minimization problem

$$\left\{\widehat{\alpha}(x), \widehat{\beta}(x)\right\} = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)(y_i - \alpha - \beta(x_i - x))^2.$$

The local linear estimator of $m(x)$ is the estimated intercept

$$\widehat{m}(x) = \widehat{\alpha}(x)$$

and the local linear estimator of the regression derivative $\nabla m(x)$ is the estimated slope coefficient

$$\widehat{\nabla m}(x) = \widehat{\beta}(x).$$

Computationally, for each x set

$$\boldsymbol{z}_i(x) = \begin{pmatrix} 1 \\ x_i - x \end{pmatrix}$$

and

$$k_i(x) = k\left(\frac{x_i - x}{h}\right).$$

Then

$$\begin{aligned}\begin{pmatrix} \widehat{\alpha}(x) \\ \widehat{\beta}(x) \end{pmatrix} &= \left(\sum_{i=1}^{n} k_i(x)\boldsymbol{z}_i(x)\boldsymbol{z}_i(x)'\right)^{-1} \sum_{i=1}^{n} k_i(x)\boldsymbol{z}_i(x)y_i \\ &= \left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{y}\end{aligned}$$

where $\boldsymbol{K} = \operatorname{diag}\{k_1(x), ..., k_n(x)\}$.

To visualize, Figure 17.2 displays the scatter plot of the same 100 observations from Figure 17.1, divided into three regions depending on the regressor x_i : $[1, 3]$, $[3, 5]$, $[5, 7]$. A linear regression is fit to the observations in each region, with the observations weighted by the Epanechnikov kernel with $h = 1$. The three fitted regression lines are displayed by the three straight solid lines. The values of these regression lines at $x = 2$, $x = 4$ and $x = 6$, respectively, are the local linear estimates $\widehat{m}(x)$ at $x = 2$, 4, and 6. This estimation is repeated for all x in the support of the regressors, and plotted as the continuous solid line in Figure 17.2.

One interesting feature is that as $h \to \infty$, the LL estimator approaches the full-sample linear least-squares estimator $\widehat{m}(x) \to \widehat{\alpha} + \widehat{\beta}x$. That is because as $h \to \infty$ all observations receive equal weight regardless of x. In this sense we can see that the LL estimator is a flexible generalization of the linear OLS estimator.

Which nonparametric estimator should you use in practice: NW or LL? The theoretical literature shows that neither strictly dominates the other, but we can describe contexts where one or the other does better. Roughly speaking, the NW estimator performs better than the LL estimator when $m(x)$ is close to a flat line, but the LL estimator performs better when $m(x)$ is meaningfully non-constant. The LL estimator also performs better for values of x near the boundary of the support of x_i.

17.5 Nonparametric Residuals and Regression Fit

The fitted regression at $x = x_i$ is $\widehat{m}(x_i)$ and the fitted residual is

$$\widehat{e}_i = y_i - \widehat{m}(x_i).$$

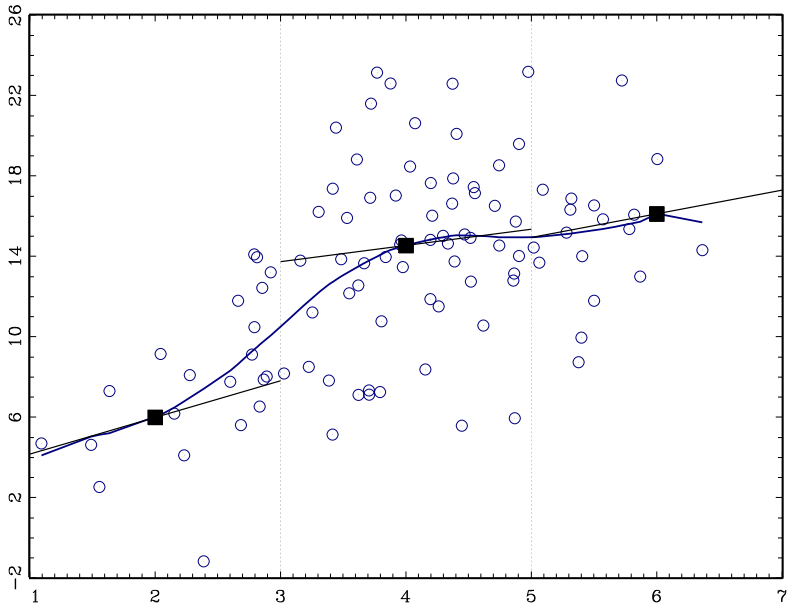


Figure 17.2: Scatter of (y_i, x_i) and Local Linear fitted regression

As a general rule, but especially when the bandwidth h is small, it is hard to view $\widehat{e}_i$ as a good measure of the fit of the regression. As $h \to 0$ then $\widehat{m}(x_i) \to y_i$ and therefore $\widehat{e}_i \to 0$. This clearly indicates overfitting as the true error is not zero. In general, since $\widehat{m}(x_i)$ is a local average which includes y_i, the fitted value will be necessarily close to y_i and the residual $\widehat{e}_i$ small, and the degree of this overfitting increases as h decreases.

A standard solution is to measure the fit of the regression at $x = x_i$ by re-estimating the model excluding the i^{th} observation. For Nadaraya-Watson regression, the leave-one-out estimator of $m(x)$ excluding observation i is

$$\widetilde{m}_{-i}(x) = \frac{\sum_{j \neq i} k\left(\dfrac{x_j - x}{h}\right) y_j}{\sum_{j \neq i} k\left(\dfrac{x_j - x}{h}\right)}.$$

Notationally, the "$-i$" subscript is used to indicate that the i^{th} observation is omitted.

The leave-one-out predicted value for y_i at $x = x_i$ equals

$$\widetilde{y}_i = \widetilde{m}_{-i}(x_i) = \frac{\sum_{j \neq i} k\left(\dfrac{x_j - x_i}{h}\right) y_j}{\sum_{j \neq i} k\left(\dfrac{x_j - x_i}{h}\right)}.$$

The leave-one-out residuals (or prediction errors) are the difference between the leave-one-out predicted values and the actual observation

$$\widetilde{e}_i = y_i - \widetilde{y}_i.$$

Since $\widetilde{y}_i$ is not a function of y_i, there is no tendency for $\widetilde{y}_i$ to overfit for small h. Consequently, $\widetilde{e}_i$ is a good measure of the fit of the estimated nonparametric regression.

Similarly, the leave-one-out local-linear residual is $\widetilde{e}_i = y_i - \widetilde{\alpha}_i$ with

$$\begin{pmatrix} \widetilde{\alpha}_i \\ \widetilde{\beta}_i \end{pmatrix} = \left(\sum_{j \neq i} k_{ij} \boldsymbol{z}_{ij} \boldsymbol{z}_{ij}' \right)^{-1} \sum_{j \neq i} k_{ij} \boldsymbol{z}_{ij} y_j,$$

$$\boldsymbol{z}_{ij} = \begin{pmatrix} 1 \\ x_j - x_i \end{pmatrix}$$

and

$$k_{ij} = k\left(\frac{x_j - x_i}{h}\right).$$

17.6 Cross-Validation Bandwidth Selection

As we mentioned before, the choice of bandwidth h is crucial. As h increases, the kernel regression estimators (both NW and LL) become more smooth, ironing out the bumps and wiggles. This reduces estimation variance but at the cost of increased bias and oversmoothing. As h decreases the estimators become more wiggly, erratic, and noisy. It is desirable to select h to trade-off these features. How can this be done systematically?

To be explicit about the dependence of the estimator on the bandwidth, let us write the estimator of $m(x)$ with a given bandwidth h as $\widehat{m}(x,h)$, and our discussion will apply equally to the NW and LL estimators.

Ideally, we would like to select h to minimize the mean-squared error (MSE) of $\widehat{m}(x,h)$ as a estimate of $m(x)$. For a given value of x the MSE is

$$MSE_n(x,h) = \mathbb{E}\left(\left(\widehat{m}(x,h) - m(x)\right)^2\right).$$

We are typically interested in estimating $m(x)$ for all values in the support of x. A common measure for the average fit is the integrated MSE

$$\begin{aligned} IMSE_n(h) &= \int MSE_n(x,h) f_x(x) dx \\ &= \int \mathbb{E}\left(\left(\widehat{m}(x,h) - m(x)\right)^2\right) f_x(x) dx \end{aligned}$$

where $f_x(x)$ is the marginal density of x_i. Notice that we have defined the IMSE as an integral with respect to the density $f_x(x)$. Other weight functions could be used, but it turns out that this is a convenient choice.

The IMSE is closely related with the MSFE of Section 4.11. Let (y_{n+1}, x_{n+1}) be out-of-sample observations (and thus independent of the sample) and consider predicting y_{n+1} given x_{n+1} and the nonparametric estimate $\widehat{m}(x,h)$. The natural point estimate for y_{n+1} is $\widehat{m}(x_{n+1},h)$ which has mean-squared forecast error

$$\begin{aligned} MSFE_n(h) &= \mathbb{E}\left(\left(y_{n+1} - \widehat{m}(x_{n+1},h)\right)^2\right) \\ &= \mathbb{E}\left(\left(e_{n+1} + m(x_{n+1}) - \widehat{m}(x_{n+1},h)\right)^2\right) \\ &= \sigma^2 + \mathbb{E}\left(\left(m(x_{n+1}) - \widehat{m}(x_{n+1},h)\right)^2\right) \\ &= \sigma^2 + \int \mathbb{E}\left(\left(\widehat{m}(x,h) - m(x)\right)^2\right) f_x(x) dx \end{aligned}$$

where the final equality uses the fact that x_{n+1} is independent of $\widehat{m}(x,h)$. We thus see that

$$MSFE_n(h) = \sigma^2 + IMSE_n(h).$$

Since σ^2 is a constant independent of the bandwidth h, $MSFE_n(h)$ and $IMSE_n(h)$ are equivalent measures of the fit of the nonparameric regression.

The optimal bandwidth h is the value which minimizes $IMSE_n(h)$ (or equivalently $MSFE_n(h)$). While these functions are unknown, we learned in Theorem 4.11.1 that (at least in the case of linear

regression) $MSFE_n$ can be estimated by the sample mean-squared prediction errors. It turns out that this fact extends to nonparametric regression. The nonparametric leave-one-out residuals are

$$\widetilde{e}_i(h) = y_i - \widetilde{m}_{-i}(x_i, h)$$

where we are being explicit about the dependence on the bandwidth h. The mean squared leave-one-out residuals is

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{e}_i(h)^2.$$

This function of h is known as the **cross-validation criterion**.

The cross-validation bandwidth $\widehat{h}$ is the value which minimizes $CV(h)$

$$\widehat{h} = \underset{h \geq h_\ell}{\operatorname{argmin}} CV(h) \tag{17.7}$$

for some $h_\ell > 0$. The restriction $h \geq h_\ell$ is imposed so that $CV(h)$ is not evaluated over unreasonably small bandwidths.

There is not an explicit solution to the minimization problem (17.7), so it must be solved numerically. A typical practical method is to create a grid of values for h, e.g. $[h_1, h_2, ..., h_J]$, evaluate $CV(h_j)$ for $j = 1, ..., J$, and set

$$\widehat{h} = \underset{h \in [h_1, h_2, ..., h_J]}{\operatorname{argmin}} CV(h).$$

Evaluation using a coarse grid is typically sufficient for practical application. Plots of $CV(h)$ against h are a useful diagnostic tool to verify that the minimum of $CV(h)$ has been obtained.

We said above that the cross-validation criterion is an estimator of the MSFE. This claim is based on the following result.

Theorem 17.6.1

$$\mathbb{E}\left(CV(h)\right) = MSFE_{n-1}(h) = IMSE_{n-1}(h) + \sigma^2 \tag{17.8}$$

Theorem 17.6.1 shows that $CV(h)$ is an unbiased estimator of $IMSE_{n-1}(h) + \sigma^2$. The first term, $IMSE_{n-1}(h)$, is the integrated MSE of the nonparametric estimator using a sample of size $n-1$. If n is large, $IMSE_{n-1}(h)$ and $IMSE_n(h)$ will be nearly identical, so $CV(h)$ is essentially unbiased as an estimator of $IMSE_n(h) + \sigma^2$. Since the second term (σ^2) is unaffected by the bandwidth h, it is irrelevant for the problem of selection of h. In this sense we can view $CV(h)$ as an estimator of the IMSE, and more importantly we can view the minimizer of $CV(h)$ as an estimate of the minimizer of $IMSE_n(h)$.

To illustrate, Figure 17.3 displays the cross-validation criteria $CV(h)$ for the Nadaraya-Watson and Local Linear estimators using the data from Figure 17.1, both using the Epanechnikov kernel. The CV functions are computed on a grid with intervals 0.01. The CV-minimizing bandwidths are $h = 1.09$ for the Nadaraya-Watson estimator and $h = 1.59$ for the local linear estimator. Figure 17.3 shows the minimizing bandwidths by the arrows. It is typical to find that the CV criteria recommends a larger bandwidth for the LL estimator than for the NW estimator, which highlights the fact that smoothing parameters such as bandwidths are specific to the particular method.

The CV criterion can also be used to select between different nonparametric estimators. The CV-selected estimator is the one with the lowest minimized CV criterion. For example, in Figure 17.3, the NW estimator has a minimized CV criterion of 16.88, while the LL estimator has a

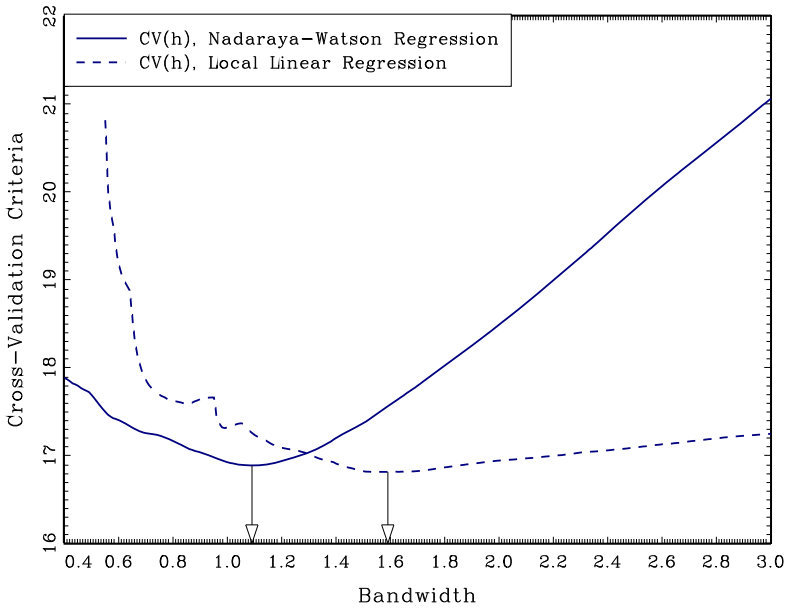


Figure 17.3: Cross-Validation Criteria, Nadaraya-Watson Regression and Local Linear Regression

minimized CV criterion of 16.81. Since the LL estimator achieves a lower value of the CV criterion, LL is the CV-selected estimator. The difference (0.07) is small, suggesting that the two estimators are near equivalent in IMSE.

Figure 17.4 displays the fitted CEF estimates (NW and LL) using the bandwidths selected by cross-validation. Also displayed is the true CEF $m(x) = 10\log(x)$. Notice that the nonparametric estimators with the CV-selected bandwidths (and especially the LL estimator) track the true CEF quite well.

Proof of Theorem 17.6.1. Observe that $m(x_i) - \widetilde{m}_{-i}(x_i, h)$ is a function only of $(x_1, ..., x_n)$ and $(e_1, ..., e_n)$ excluding e_i, and is thus uncorrelated with e_i. Since $\widetilde{e}_i(h) = m(x_i) - \widetilde{m}_{-i}(x_i, h) + e_i$, then

$$\begin{aligned}
\mathbb{E}\left(CV(h)\right) &= \mathbb{E}\left(\widetilde{e}_i(h)^2\right) \\
&= \mathbb{E}\left(e_i^2\right) + \mathbb{E}\left(\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right)^2\right) \\
&+ 2\mathbb{E}\left(\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right) e_i\right) \\
&= \sigma^2 + \mathbb{E}\left(\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right)^2\right). \qquad (17.9)
\end{aligned}$$

The second term is an expectation over the random variables x_i and $\widetilde{m}_{-i}(x, h)$, which are independent as the second is not a function of the i^{th} observation. Thus taking the conditional expectation given the sample excluding the i^{th} observation, this is the expectation over x_i only, which is the integral with respect to its density

$$\mathbb{E}_{-i}\left(\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right)^2\right) = \int \left(\widetilde{m}_{-i}(x, h) - m(x)\right)^2 f_x(x)dx.$$

Taking the unconditional expecation yields

$$\begin{aligned}
\mathbb{E}\left(\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right)^2\right) &= \mathbb{E}\int \left(\widetilde{m}_{-i}(x, h) - m(x)\right)^2 f_x(x)dx \\
&= IMSE_{n-1}(h)
\end{aligned}$$

where this is the IMSE of a sample of size $n-1$ as the estimator $\widetilde{m}_{-i}$ uses $n-1$ observations. Combined with (17.9) we obtain (17.8), as desired. ■

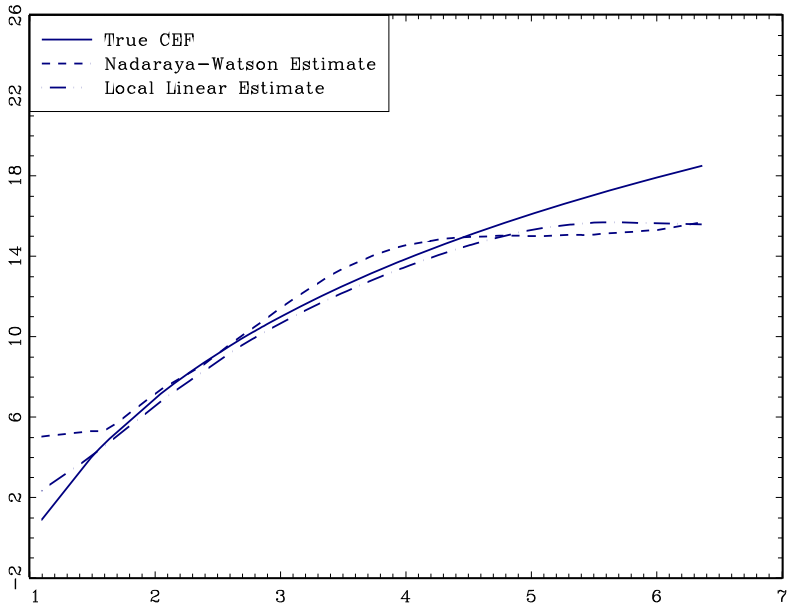


Figure 17.4: Nonparametric Estimates using data-dependent (CV) bandwidths

17.7 Asymptotic Distribution

There is no finite sample distribution theory for kernel estimators, but there is a well developed asymptotic distribution theory. The theory is based on the approximation that the bandwidth h decreases to zero as the sample size n increases. This means that the smoothing is increasingly localized as the sample size increases. So long as the bandwidth does not decrease to zero too quickly, the estimator can be shown to be asymptotically normal, but with a non-trivial bias.

Let $f_x(x)$ denote the marginal density of x_i and $\sigma^2(x) = \mathbb{E}\left(e_i^2 \mid x_i = x\right)$ denote the conditional variance of $e_i = y_i - m(x_i)$.

Theorem 17.7.1 *Let $\widehat{m}(x)$ denote either the Nadarya-Watson or Local Linear estimator of $m(x)$. If x is interior to the support of x_i and $f_x(x) > 0$, then as $n \to \infty$ and $h \to 0$ such that $nh \to \infty$,*

$$\sqrt{nh}\left(\widehat{m}(x) - m(x) - h^2\sigma_k^2 B(x)\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{R_k \sigma^2(x)}{f_x(x)}\right) \tag{17.10}$$

where σ_k^2 and R_k are defined in (17.5) and (17.6). For the Nadaraya-Watson estimator

$$B(x) = \frac{1}{2}m''(x) + f_x(x)^{-1} f_x'(x) m'(x)$$

and for the local linear estimator

$$B(x) = \frac{1}{2}f_x(x) m''(x)$$

There are several interesting features about the asymptotic distribution which are noticeably different than for parametric estimators. First, the estimator converges at the rate $\sqrt{nh}$, not $\sqrt{n}$.

Since $h \to 0$, $\sqrt{nh}$ diverges slower than $\sqrt{n}$, thus the nonparametric estimator converges more slowly than a parametric estimator. Second, the asymptotic distribution contains a non-neglible bias term $h^2\sigma_k^2 B(x)$. This term asymptotically disappears since $h \to 0$. Third, the assumptions that $nh \to \infty$ and $h \to 0$ mean that the estimator is consistent for the CEF $m(x)$.

The fact that the estimator converges at the rate $\sqrt{nh}$ has led to the interpretation of nh as the "effective sample size". This is because the number of observations being used to construct $\widehat{m}(x)$ is proportional to nh, not n as for a parametric estimator.

It is helpful to understand that the nonparametric estimator has a reduced convergence rate because the object being estimated – $m(x)$ – is nonparametric. This is harder than estimating a finite dimensional parameter, and thus comes at a cost.

Unlike parametric estimation, the asymptotic distribution of the nonparametric estimator includes a term representing the bias of the estimator. The asymptotic distribution (17.10) shows the form of this bias. Not only is it proportional to the squared bandwidth h^2 (the degree of smoothing), it is proportional to the function $B(x)$ which depends on the slope and curvature of the CEF $m(x)$. Interestingly, when $m(x)$ is constant then $B(x) = 0$ and the kernel estimator has no asymptotic bias. The bias is essentially increasing in the curvature of the CEF function $m(x)$. This is because the local averaging smooths $m(x)$, and the smoothing induces more bias when $m(x)$ is curved.

Theorem 17.7.1 shows that the asymptotic distributions of the NW and LL estimators are similar, with the only difference arising in the bias function $B(x)$. The bias term for the NW estimator has an extra component which depends on the first derivative of the CEF $m(x)$ while the bias term of the LL estimator is invariant to the first derivative. The fact that the bias formula for the LL estimator is simpler and is free of dependence on the first derivative of $m(x)$ suggests that the LL estimator will generally have smaller bias than the NW estimator (but this is not a precise ranking). Since the asymptotic variances in the two distributions are the same, this means that the LL estimator achieves a reduced bias without an effect on asymptotic variance. This analysis has led to the general preference for the LL estimator over the NW estimator in the nonparametrics literature.

One implication of Theorem 17.7.1 is that we can define the asymptotic MSE (AMSE) of $\widehat{m}(x)$ as the squared bias plus the asymptotic variance

$$AMSE\left(\widehat{m}(x)\right) = \left(h^2\sigma_k^2 B(x)\right)^2 + \frac{R_k \sigma^2(x)}{nh f_x(x)}. \tag{17.11}$$

Focusing on rates, this says

$$AMSE\left(\widehat{m}(x)\right) \sim h^4 + \frac{1}{nh} \tag{17.12}$$

which means that the AMSE is dominated by the larger of h^4 and $(nh)^{-1}$. Notice that the bias is increasing in h and the variance is decreasing in h. (More smoothing means more observations are used for local estimation: this increases the bias but decreases estimation variance.) To select h to minimize the AMSE, these two components should balance each other. Setting $h^4 \propto (nh)^{-1}$ means setting $h \propto n^{-1/5}$. Another way to see this is to pick h to minimize the right-hand-side of (17.12). The first-order condition for h is

$$\frac{\partial}{\partial h}\left(h^4 + \frac{1}{nh}\right) = 4h^3 - \frac{1}{nh^2} = 0$$

which when solved for h yields $h = n^{-1/5}$. What this means is that for AMSE-efficient estimation of $m(x)$, the optimal rate for the bandwidth is $h \propto n^{-1/5}$.

Theorem 17.7.2 *The bandwidth which minimizes the AMSE (17.12) is of order $h \propto n^{-1/5}$. With $h \propto n^{-1/5}$ then $AMSE\left(\widehat{m}(x)\right) = O\left(n^{-4/5}\right)$ and $\widehat{m}(x) = m(x) + O_p\left(n^{-2/5}\right)$.*

This result means that the bandwidth should take the form $h = cn^{-1/5}$. The optimal constant c depends on the kernel k, the bias function $B(x)$ and the marginal density $f_x(x)$. A common misinterpretation is to set $h = n^{-1/5}$, which is equivalent to setting $c = 1$ and is completely arbitrary. Instead, an empirical bandwidth selection rule such as cross-validation should be used in practice.

When $h = cn^{-1/5}$ we can rewrite the asymptotic distribution (17.10) as

$$n^{2/5}\left(\widehat{m}(x) - m(x)\right) \xrightarrow{d} \mathrm{N}\left(c^2\sigma_k^2 B(x), \frac{R_k\sigma^2(x)}{cf_x(x)}\right)$$

In this representation, we see that $\widehat{m}(x)$ is asymptotically normal, but with a $n^{2/5}$ rate of convergence and non-zero mean. The asymptotic distribution depends on the constant c through the bias (positively) and the variance (inversely).

The asymptotic distribution in Theorem 17.7.1 allows for the optimal rate $h = cn^{-1/5}$ but this rate is not required. In particular, consider an undersmoothing (smaller than optimal) bandwith with rate $h = o\left(n^{-1/5}\right)$. For example, we could specify that $h = cn^{-\alpha}$ for some $c > 0$ and $1/5 < \alpha < 1$. Then $\sqrt{nh}h^2 = O(n^{(1-5\alpha)/2}) = o(1)$ so the bias term in (17.10) is asymptotically negligible so Theorem 17.7.1 implies

$$\sqrt{nh}\left(\widehat{m}(x) - m(x)\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{R_k\sigma^2(x)}{f_x(x)}\right).$$

That is, the estimator is asymptotically normal without a bias component. Not having an asymptotic bias component is convenient for some theoretical manipuations, so many authors impose the undersmoothing condition $h = o\left(n^{-1/5}\right)$ to ensure this situation. This convenience comes at a cost. First, the resulting estimator is inefficient as its convergence rate is is $O_p\left(n^{-(1-\alpha)/2}\right) > O_p\left(n^{-2/5}\right)$ since $\alpha > 1/5$. Second, the distribution theory is an inherently misleading approximation as it misses a critically key ingredient of nonparametric estimation – the trade-off between bias and variance. The approximation (17.10) is superior precisely because it contains the asymptotic bias component which is a realistic implication of nonparametric estimation. Undersmoothing assumptions should be avoided when possible.

17.8 Conditional Variance Estimation

Let's consider the problem of estimation of the conditional variance

$$\begin{aligned}\sigma^2(x) &= \mathrm{var}\left(y_i \mid x_i = x\right) \\ &= \mathbb{E}\left(e_i^2 \mid x_i = x\right).\end{aligned}$$

Even if the conditional mean $m(x)$ is parametrically specified, it is natural to view $\sigma^2(x)$ as inherently nonparametric as economic models rarely specify the form of the conditional variance. Thus it is quite appropriate to estimate $\sigma^2(x)$ nonparametrically.

We know that $\sigma^2(x)$ is the CEF of e_i^2 given x_i. Therefore if e_i^2 were observed, $\sigma^2(x)$ could be nonparametrically estimated using NW or LL regression. For example, the ideal NW estimator is

$$\overline{\sigma}^2(x) = \frac{\sum_{i=1}^n k_i(x)e_i^2}{\sum_{i=1}^n k_i(x)}.$$

Since the errors e_i are not observed, we need to replace them with an empirical residual, such as $\widehat{e}_i = y_i - \widehat{m}(x_i)$ where $\widehat{m}(x)$ is the estimated CEF. (The latter could be a nonparametric estimator such as NW or LL, or even a parametric estimator.) Even better, use the leave-one-out prediction errors $\widetilde{e}_i = y_i - \widehat{m}_{-i}(x_i)$, as these are not subject to overfitting.

With this substitution the NW estimator of the conditional variance is

$$\widehat{\sigma}^2(x) = \frac{\sum_{i=1}^n k_i(x)\widetilde{e}_i^2}{\sum_{i=1}^n k_i(x)}. \tag{17.13}$$

This estimator depends on a set of bandwidths $h_1, ..., h_q$, but there is no reason for the bandwidths to be the same as those used to estimate the conditional mean. Cross-validation can be used to select the bandwidths for estimation of $\widehat{\sigma}^2(x)$ separately from cross-validation for estimation of $\widehat{m}(x)$.

There is one subtle difference between CEF and conditional variance estimation. The conditional variance is inherently non-negative $\sigma^2(x) \geq 0$ and it is desirable for our estimator to satisfy this property. Interestingly, the NW estimator (17.13) is necessarily non-negative, since it is a smoothed average of the non-negative squared residuals, but the LL estimator is not guarenteed to be non-negative for all x. For this reason, the NW estimator is preferred for conditional variance estimation.

Fan and Yao (1998, Biometrika) derive the asymptotic distribution of the estimator (17.13). They obtain the surprising result that the asymptotic distribution of this two-step estimator is identical to that of the one-step idealized estimator $\overline{\sigma}^2(x)$.

17.9 Standard Errors

Theorem 17.7.1 shows the asymptotic variances of both the NW and LL nonparametric regression estimators equal

$$V(x) = \frac{R_k \sigma^2(x)}{f_x(x)}.$$

For standard errors we need an estimate of $V(x)$. A plug-in estimate replaces the unknowns by estimates. The roughness R_k can be found from Table 9.1. The conditional variance can be estimated using (17.13). The density of x_i can be estimated using the methods from Section 22.1. Replacing these estimates into the formula for $V(x)$ we obtain the asymptotic variance estimate

$$\widehat{V}(x) = \frac{R_k \widehat{\sigma}^2(x)}{\widehat{f}_x(x)}.$$

Then an asymptotic standard error for the kernel estimate $\widehat{m}(\boldsymbol{x})$ is

$$\widehat{s}(x) = \sqrt{\frac{1}{nh}\widehat{V}(x)}.$$

Plots of the estimated CEF $\widehat{m}(x)$ can be accompanied by confidence intervals $\widehat{m}(x) \pm 2\widehat{s}(x)$. These are known as **pointwise confidence intervals**, as they are designed to have correct coverage at each x, not uniformly in x.

One important caveat about the interpretation of nonparametric confidence intervals is that they are not centered at the true CEF $m(x)$, but rather are centered at the biased or pseudo-true value

$$m^*(x) = m(x) + h^2 \sigma_k^2 B(x).$$

Consequently, a correct statement about the confidence interval $\widehat{m}(x) \pm 2\widehat{s}(x)$ is that it asymptotically contains $m^*(x)$ with probability 95%, not that it asymptotically contains $m(x)$ with probability 95%. The discrepancy is that the confidence interval does not take into account the bias $h^2\sigma_k^2 B(x)$. Unfortunately, nothing constructive can be done about this. The bias is difficult and noisy to estimate, so making a bias-correction only inflates estimation variance and decreases overall precision. A technical "trick" is to assume undersmoothing $h = o\left(n^{-1/5}\right)$ but this does not really eliminate the bias, it only assumes it away. The plain fact is that once we honestly acknowledge that the true CEF is nonparametric, it then follows that any finite sample estimate will have finite sample bias, and this bias will be inherently unknown and thus impossible to incorporate into confidence intervals.

17.10 Multiple Regressors

Our analysis has focus on the case of real-valued x_i for simplicity of exposition, but the methods of kernel regression extend easily to the multiple regressor case, at the cost of a reduced rate of convergence. In this section we consider the case of estimation of the conditional expectation function

$$\mathbb{E}\left(y_i \mid \boldsymbol{x}_i = \boldsymbol{x}\right) = m(\boldsymbol{x})$$

when

$$\boldsymbol{x}_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{di} \end{pmatrix}$$

is a d-vector.

For any evaluation point $\boldsymbol{x}$ and observation i, define the kernel weights

$$k_i(\boldsymbol{x}) = k\left(\frac{x_{1i} - x_1}{h_1}\right) k\left(\frac{x_{2i} - x_2}{h_2}\right) \cdots k\left(\frac{x_{di} - x_d}{h_d}\right),$$

a d-fold product kernel. The kernel weights $k_i(\boldsymbol{x})$ assess if the regressor vector $\boldsymbol{x}_i$ is close to the evaluation point $\boldsymbol{x}$ in the Euclidean space $\mathbb{R}^d$.

These weights depend on a set of d bandwidths, h_j, one for each regressor. We can group them together into a single vector for notational convenience:

$$\boldsymbol{h} = \begin{pmatrix} h_1 \\ \vdots \\ h_d \end{pmatrix}.$$

Given these weights, the Nadaraya-Watson estimator takes the form

$$\widehat{m}(\boldsymbol{x}) = \frac{\sum_{i=1}^n k_i(\boldsymbol{x}) y_i}{\sum_{i=1}^n k_i(\boldsymbol{x})}.$$

For the local-linear estimator, define

$$\boldsymbol{z}_i(\boldsymbol{x}) = \begin{pmatrix} 1 \\ \boldsymbol{x}_i - \boldsymbol{x} \end{pmatrix}$$

and then the local-linear estimator can be written as $\widehat{m}(\boldsymbol{x}) = \widehat{\alpha}(\boldsymbol{x})$ where

$$\begin{aligned} \begin{pmatrix} \widehat{\alpha}(\boldsymbol{x}) \\ \widehat{\beta}(\boldsymbol{x}) \end{pmatrix} &= \left(\sum_{i=1}^n k_i(\boldsymbol{x}) \boldsymbol{z}_i(\boldsymbol{x}) \boldsymbol{z}_i(\boldsymbol{x})'\right)^{-1} \sum_{i=1}^n k_i(\boldsymbol{x}) \boldsymbol{z}_i(\boldsymbol{x}) y_i \\ &= \left(\boldsymbol{Z}'\boldsymbol{K}\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\boldsymbol{K}\boldsymbol{y} \end{aligned}$$

where $\boldsymbol{K} = \operatorname{diag}\{k_1(x), ..., k_n(x)\}$.

In multiple regressor kernel regression, cross-validation remains a recommended method for bandwidth selection. The leave-one-out residuals $\widetilde{e}_i$ and cross-validation criterion $CV(\boldsymbol{h})$ are defined identically as in the single regressor case. The only difference is that now the CV criterion is a function over the d-dimensional bandwidth $\boldsymbol{h}$. This is a critical practical difference since finding the bandwidth vector $\widehat{\boldsymbol{h}}$ which minimizes $CV(\boldsymbol{h})$ can be computationally difficult when $\boldsymbol{h}$ is high dimensional. Grid search is cumbersome and costly, since G gridpoints per dimension imply evaulation of $CV(\boldsymbol{h})$ at G^d distinct points, which can be a large number. Furthermore, plots of $CV(\boldsymbol{h})$ against $\boldsymbol{h}$ are challenging when $d > 2$.

The asymptotic distribution of the estimators in the multiple regressor case is an extension of the single regressor case. Let $f_x(\boldsymbol{x})$ denote the marginal density of $\boldsymbol{x}_i$ and $\sigma^2(\boldsymbol{x}) = \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i = \boldsymbol{x}\right)$ the conditional variance of $e_i = y_i - m(\boldsymbol{x}_i)$. Let $|\boldsymbol{h}| = h_1 h_2 \cdots h_d$.

Theorem 17.10.1 *Let $\widehat{m}(\boldsymbol{x})$ denote either the Nadarya-Watson or Local Linear estimator of $m(\boldsymbol{x})$. If $\boldsymbol{x}$ is interior to the support of $\boldsymbol{x}_i$ and $f_x(\boldsymbol{x}) > 0$, then as $n \to \infty$ and $h_j \to 0$ such that $n\,|\boldsymbol{h}| \to \infty$,*

$$\sqrt{n\,|\boldsymbol{h}|}\left(\widehat{m}(\boldsymbol{x}) - m(\boldsymbol{x}) - \sigma_k^2 \sum_{j=1}^{d} h_j^2 B_{,j}(\boldsymbol{x})\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{R_k^d \sigma^2(\boldsymbol{x})}{f_x(\boldsymbol{x})}\right)$$

where for the Nadaraya-Watson estimator

$$B_j(\boldsymbol{x}) = \frac{1}{2}\frac{\partial^2}{\partial x_j^2} m(\boldsymbol{x}) + f_x(\boldsymbol{x})^{-1} \frac{\partial}{\partial x_j} f_x(\boldsymbol{x}) \frac{\partial}{\partial x_j} m(\boldsymbol{x})$$

and for the Local Linear estimator

$$B_j(\boldsymbol{x}) = \frac{1}{2}\frac{\partial^2}{\partial x_j^2} m(\boldsymbol{x})$$

For notational simplicity consider the case that there is a single common bandwidth h. In this case the AMSE takes the form

$$AMSE(\widehat{m}(\boldsymbol{x})) \sim h^4 + \frac{1}{nh^d}$$

That is, the squared bias is of order h^4, the same as in the single regressor case, but the variance is of larger order $(nh^d)^{-1}$. Setting h to balance these two components requires setting $h \sim n^{-1/(4+d)}$.

Theorem 17.10.2 *The bandwidth which minimizes the AMSE is of order $h \propto n^{-1/(4+d)}$. With $h \propto n^{-1/(4+d)}$ then $AMSE\left(\widehat{m}(\boldsymbol{x})\right) = O\left(n^{-4/(4+d)}\right)$ and $\widehat{m}(\boldsymbol{x}) = m(\boldsymbol{x}) + O_p\left(n^{-2/(4+d)}\right)$*

In all estimation problems an increase in the dimension decreases estimation precision. For example, in parametric estimation an increase in dimension typically increases the asymptotic variance. In nonparametric estimation an increase in the dimension typically decreases the convergence rate, which is a more fundamental decrease in precision. For example, in kernel regression the convergence rate $O_p\left(n^{-2/(4+d)}\right)$ decreases as d increases. The reason is the estimator $\widehat{m}(\boldsymbol{x})$ is a local average of the y_i for observations such that $\boldsymbol{x}_i$ is close to $\boldsymbol{x}$, and when there are multiple regressors the number of such observations is inherently smaller. This phenomenon – that the rate of convergence of nonparametric estimation decreases as the dimension increases – is called the **curse of dimensionality**.

Chapter 18

Series Estimation

18.1 Approximation by Series

As we mentioned at the beginning of Chapter 17, there are two main methods of nonparametric regression: kernel estimation and series estimation. In this chapter we study series methods.

Series methods approximate an unknown function (e.g. the CEF $m(\boldsymbol{x})$) with a flexible parametric function, with the number of parameters treated similarly to the bandwidth in kernel regression. A series approximation to $m(\boldsymbol{x})$ takes the form $m_K(\boldsymbol{x}) = m_K(\boldsymbol{x}, \boldsymbol{\beta}_K)$ where $m_K(\boldsymbol{x}, \boldsymbol{\beta}_K)$ is a known parametric family and $\boldsymbol{\beta}_K$ is an unknown coefficient. The integer K is the dimension of $\boldsymbol{\beta}_K$ and indexes the complexity of the approximation.

A linear series approximation takes the form

$$\begin{aligned} m_K(\boldsymbol{x}) &= \sum_{j=1}^{K} z_{jK}(\boldsymbol{x})\beta_{jK} \\ &= \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K \end{aligned} \tag{18.1}$$

where $z_{jK}(\boldsymbol{x})$ are (nonlinear) functions of $\boldsymbol{x}$, and are known as **basis functions** or **basis function transformations** of $\boldsymbol{x}$.

For real-valued x, a well-known linear series approximation is the p^{th}-order **polynomial**

$$m_K(x) = \sum_{j=0}^{p} x^j \beta_{jK}$$

where $K = p + 1$.

When $\boldsymbol{x} \in \mathbb{R}^d$ is vector-valued, a p^{th}-order polynomial is

$$m_K(\boldsymbol{x}) = \sum_{j_1=0}^{p} \cdots \sum_{j_d=0}^{p} x_1^{j_1} \cdots x_d^{j_d} \beta_{j_1,\ldots,j_d K}.$$

This includes all powers and cross-products, and the coefficient vector has dimension $K = (p+1)^d$. In general, a common method to create a series approximation for vector-valued $\boldsymbol{x}$ is to include all non-redundant cross-products of the basis function transformations of the components of $\boldsymbol{x}$.

18.2 Splines

Another common series approximation is a continuous piecewise polynomial function known as a **spline**. While splines can be of any polynomial order (e.g. linear, quadratic, cubic, etc.), a common choice is cubic. To impose smoothness it is common to constrain the spline function to have continuous derivatives up to the order of the spline. Thus a quadratic spline is typically

constrained to have a continuous first derivative, and a cubic spline is typically constrained to have a continuous first and second derivative.

There is more than one way to define a spline series expansion. All are based on the number of **knots** – the join points between the polynomial segments.

To illustrate, a piecewise linear function with two segments and a knot at t is

$$m_K(x) = \begin{cases} m_1(x) = \beta_{00} + \beta_{01}(x - t) & x < t \\ \\ m_2(x) = \beta_{10} + \beta_{11}(x - t) & x \geq t \end{cases}$$

(For convenience we have written the segments functions as polyomials in $x - t$.) The function $m_K(x)$ equals the linear function $m_1(x)$ for $x < t$ and equals $m_2(t)$ for $x > t$. Its left limit at $x = t$ is β_{00} and its right limit is β_{10}, so is continuous if (and only if) $\beta_{00} = \beta_{10}$. Enforcing this constraint is equivalent to writing the function as

$$m_K(x) = \beta_0 + \beta_1(x - t) + \beta_2(x - t)\,1(x \geq t)$$

or after transforming coefficients, as

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2(x - t)\,1(x \geq t).$$

Notice that this function has $K = 3$ coefficients, the same as a quadratic polynomial.

A piecewise quadratic function with one knot at t is

$$m_K(x) = \begin{cases} m_1(x) = \beta_{00} + \beta_{01}(x - t) + \beta_{02}(x - t)^2 & x < t \\ \\ m_2(x) = \beta_{10} + \beta_{11}(x - t) + \beta_{12}(x - t)^2 & x \geq t \end{cases}$$

This function is continuous at $x = t$ if $\beta_{00} = \beta_{10}$, and has a continuous first derivative if $\beta_{01} = \beta_{11}$. Imposing these contraints and rewriting, we obtain the function

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3(x - t)^2\,1(x \geq t).$$

Here, $K = 4$.

Furthermore, a piecewise cubic function with one knot and a continuous second derivative is

$$m_K(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x - t)^3\,1(x \geq t)$$

which has $K = 5$.

The polynomial order p is selected to control the smoothness of the spline, as $m_K(x)$ has continuous derivatives up to $p - 1$.

In general, a p^{th}-order spline with N knots at t_1, t_2, ..., t_N with $t_1 < t_2 < \cdots < t_N$ is

$$m_K(x) = \sum_{j=0}^{p} \beta_j x^j + \sum_{k=1}^{N} \gamma_k (x - t_k)^p\,1(x \geq t_k)$$

which has $K = N + p + 1$ coefficients.

In spline approximation, the typical approach is to treat the polynomial order p as fixed, and select the number of knots N to determine the complexity of the approximation. The knots t_k are typically treated as fixed. A common choice is to set the knots to evenly partition the support $\mathcal{X}$ of $\boldsymbol{x}_i$.

18.3 Partially Linear Model

A common use of a series expansion is to allow the CEF to be nonparametric with respect to one variable, yet linear in the other variables. This allows flexibility in a particular variable of interest. A partially linear CEF with vector-valued regressor $\boldsymbol{x}_1$ and real-valued continuous x_2 takes the form

$$m\left(\boldsymbol{x}_1, x_2\right) = \boldsymbol{x}_1'\boldsymbol{\beta}_1 + m_2(x_2).$$

This model is commonly used when $\boldsymbol{x}_1$ are discrete (e.g. binary variables) and x_2 is continuously distributed.

Series methods are particularly convenient for estimation of partially linear models, as we can replace the unknown function $m_2(x_2)$ with a series expansion to obtain

$$\begin{aligned} m\left(\boldsymbol{x}\right) &\simeq m_K\left(\boldsymbol{x}\right) \\ &= \boldsymbol{x}_1'\boldsymbol{\beta}_1 + \boldsymbol{z}_K'\boldsymbol{\beta}_{2K} \\ &= \boldsymbol{x}_K'\boldsymbol{\beta}_K \end{aligned}$$

where $\boldsymbol{z}_K = \boldsymbol{z}_K(x_2)$ are the basis transformations of x_2 (typically polynomials or splines) and $\boldsymbol{\beta}_{2K}$ are coefficients. After transformation the regressors are $\boldsymbol{x}_K = (\boldsymbol{x}_1', \boldsymbol{z}_K')$. and the coefficients are $\boldsymbol{\beta}_K = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_{2K}')'$.

18.4 Additively Separable Models

When $\boldsymbol{x}$ is multivariate a common simplification is to treat the regression function $m\left(\boldsymbol{x}\right)$ as additively separable in the individual regressors, which means that

$$m\left(\boldsymbol{x}\right) = m_1\left(x_1\right) + m_2\left(x_2\right) + \cdots + m_d\left(x_d\right).$$

Series methods are quite convenient for additively separable models, as we simply apply series expansions (polynomials or splines) separately for each component $m_j\left(x_j\right)$. The advantage of additive separability is the reduction in dimensionality. While an unconstrained p^{th} order polynomial has $(p+1)^d$ coefficients, an additively separable polynomial model has only $(p+1)d$ coefficients. This can be a major reduction in the number of coefficients. The disadvantage of this simplification is that the interaction effects have been eliminated.

The decision to impose additive separability can be based on an economic model which suggests the absence of interaction effects, or can be a model selection decision similar to the selection of the number of series terms. We will discuss model selection methods below.

18.5 Uniform Approximations

A good series approximation $m_K(\boldsymbol{x})$ will have the property that it gets close to the true CEF $m(\boldsymbol{x})$ as the complexity K increases. Formal statements can be derived from the theory of functional analysis.

An elegant and famous theorem is the **Stone-Weierstrass theorem**, (Weierstrass, 1885, Stone 1937, 1948) which states that any continuous function can be arbitrarily uniformly well approximated by a polynomial of sufficiently high order. Specifically, the theorem states that for $\boldsymbol{x} \in \mathbb{R}^d$, if $m(\boldsymbol{x})$ is continuous on a compact set $\mathcal{X}$, then for any $\varepsilon > 0$ there exists a polynomial $m_K(\boldsymbol{x})$ of some order K which is uniformly within ε of $m(\boldsymbol{x})$:

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left|m_K(\boldsymbol{x}) - m(\boldsymbol{x})\right| \leq \varepsilon. \tag{18.2}$$

Thus the true unknown $m(\boldsymbol{x})$ can be arbitrarily well approximately by selecting a suitable polynomial.

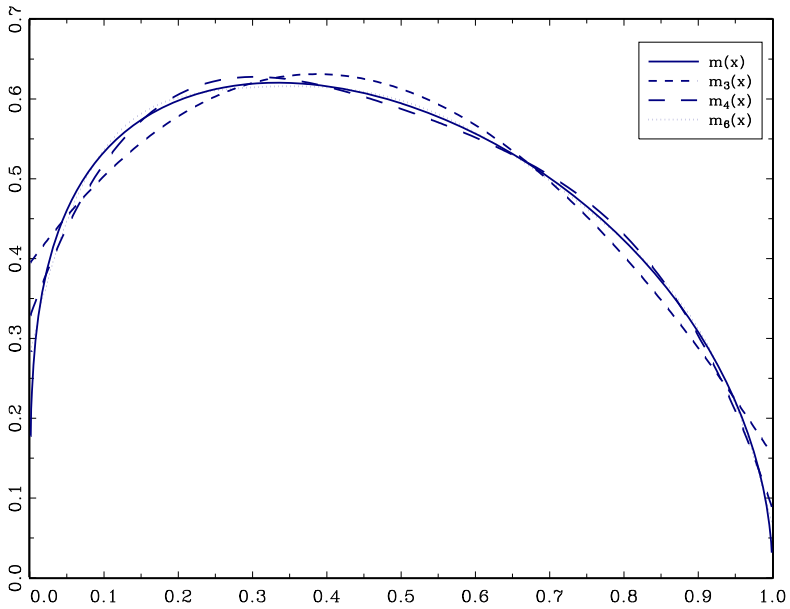


Figure 18.1: True CEF and Best Approximations

The result (18.2) can be stengthened. In particular, if the s^{th} derivative of $m(\boldsymbol{x})$ is continuous then the uniform approximation error satisfies

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |m_K(\boldsymbol{x}) - m(\boldsymbol{x})| = O\left(K^{-\alpha}\right) \tag{18.3}$$

as $K \to \infty$ where $\alpha = s/d$. This result is more useful than (18.2) because it gives a rate at which the approximation $m_K(\boldsymbol{x})$ approaches $m(\boldsymbol{x})$ as K increases.

Both (18.2) and (18.3) hold for spline approximations as well.

Intuitively, the number of derivatives s indexes the smoothness of the function $m(\boldsymbol{x})$. (18.3) says that the best rate at which a polynomial or spline approximates the CEF $m(\boldsymbol{x})$ depends on the underlying smoothness of $m(\boldsymbol{x})$. The more smooth is $m(\boldsymbol{x})$, the fewer series terms (polynomial order or spline knots) are needed to obtain a good approximation.

To illustrate polynomial approximation, Figure 18.1 displays the CEF $m(x) = x^{1/4}(1-x)^{1/2}$ on $x \in [0,1]$. In addition, the best approximations using polynomials of order $K = 3$, $K = 4$, and $K = 6$ are displayed. You can see how the approximation with $K = 3$ is fairly crude, but improves with $K = 4$ and especially $K = 6$. Approximations obtained with cubic splines are quite similar so not displayed.

As a series approximation can be written as $m_K(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K$ as in (18.1), then the coefficient of the best uniform approximation (18.3) is then

$$\boldsymbol{\beta}_K^* = \underset{\boldsymbol{\beta}_K}{\operatorname{argmin}} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K - m(\boldsymbol{x})\right|. \tag{18.4}$$

The approximation error is

$$r_K^*(\boldsymbol{x}) = m(\boldsymbol{x}) - \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K^*.$$

We can write this as

$$m(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K^* + r_K^*(\boldsymbol{x}) \tag{18.5}$$

to emphasize that the true conditional mean can be written as the linear approximation plus error. A useful consequence of equation (18.3) is

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |r_K^*(\boldsymbol{x})| \le O\left(K^{-\alpha}\right). \tag{18.6}$$

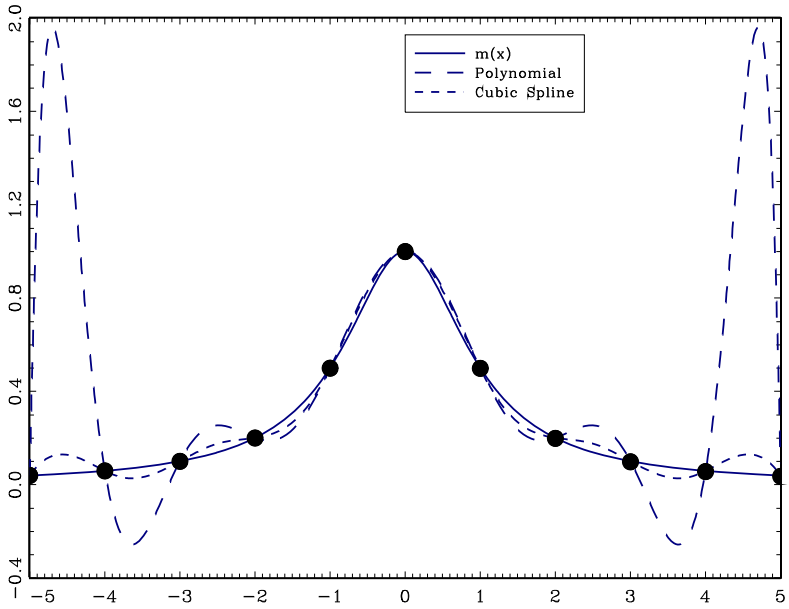


Figure 18.2: True CEF, polynomial interpolation, and spline interpolation

18.6 Runge's Phenomenon

Despite the excellent approximation implied by the Stone-Weierstrass theorem, polynomials have the troubling disadvantage that they are very poor at simple interpolation. The problem is known as **Runge's phenomenon**, and is illustrated in Figure 18.2. The solid line is the CEF $m(x) = (1 + x^2)^{-1}$ displayed on $[-5, 5]$. The circles display the function at the $K = 11$ integers in this interval. The long dashes display the 10^{th} order polynomial fit through these points. Notice that the polynomial approximation is erratic and far from the smooth CEF. This discrepancy gets worse as the number of evaluation points increases, as Runge (1901) showed that the discrepancy increases to infinity with K.

In contrast, splines do not exhibit Runge's phenomenon. In Figure 18.2 the short dashes display a cubic spline with seven knots fit through the same points as the polynomial. While the fitted spline displays some oscillation relative to the true CEF, they are relatively moderate.

Because of Runge's phenomenon, high-order polynomials are not used for interpolation, and are not popular choices for high-order series approximations. Instead, splines are widely used.

18.7 Approximating Regression

For each observation i we observe $(y_i, \boldsymbol{x}_i)$ and then construct the regressor vector $\boldsymbol{z}_{Ki} = \boldsymbol{z}_K(\boldsymbol{x}_i)$ using the series transformations. Stacking the observations in the matrices $\boldsymbol{y}$ and $\boldsymbol{Z}_K$, the least squares estimate of the coefficient $\boldsymbol{\beta}_K$ in the series approximation $\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K$ is

$$\widehat{\boldsymbol{\beta}}_K = \left(\boldsymbol{Z}_K'\boldsymbol{Z}_K\right)^{-1}\boldsymbol{Z}_K'\boldsymbol{y},$$

and the least squares estimate of the regression function is

$$\widehat{m}_K(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})'\widehat{\boldsymbol{\beta}}_K. \tag{18.7}$$

As we learned in Chapter 2, the least-squares coefficient is estimating the best linear predictor of y_i given $\boldsymbol{z}_{Ki}$. This is

$$\boldsymbol{\beta}_K = \mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}y_i\right).$$

Given this coefficient, the series approximation is $\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K$ with approximation error

$$r_K(\boldsymbol{x}) = m(\boldsymbol{x}) - \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K. \tag{18.8}$$

The true CEF equation for y_i is

$$y_i = m(\boldsymbol{x}_i) + e_i \tag{18.9}$$

with e_i the CEF error. Defining $r_{Ki} = r_K(\boldsymbol{x}_i)$, we find

$$y_i = \boldsymbol{z}_{Ki}'\boldsymbol{\beta}_K + e_{Ki}$$

where the equation error is

$$e_{Ki} = r_{Ki} + e_i.$$

Observe that the error e_{Ki} includes the approximation error and thus does not have the properties of a CEF error.

In matrix notation we can write these equations as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Z}_K\boldsymbol{\beta}_K + \boldsymbol{r}_K + \boldsymbol{e} \\ &= \boldsymbol{Z}_K\boldsymbol{\beta}_K + \boldsymbol{e}_K. \end{aligned} \tag{18.10}$$

We now impose some regularity conditions on the regression model to facilitate the theory.

Define the $K \times K$ expected design matrix

$$\boldsymbol{Q}_K = \mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right),$$

let $\mathcal{X}$ denote the support of $\boldsymbol{x}_i$, and define the largest normalized length of the regressor vector in the support of $\boldsymbol{x}_i$

$$\zeta_K = \sup_{\boldsymbol{x}\in\mathcal{X}} \left(\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{Q}_K^{-1}\boldsymbol{z}_K(\boldsymbol{x})\right)^{1/2}. \tag{18.11}$$

$\boldsymbol{\zeta}_K$ will increase with K. For example, if the support of the variables $\boldsymbol{z}_K(\boldsymbol{x}_i)$ is the unit cube $[0,1]^K$, then you can compute that $\zeta_K = \sqrt{K}$. As discussed in Newey (1997) and Li and Racine (2007, Corollary 15.1) if the support of $\boldsymbol{x}_i$ is compact then $\zeta_K = O(K)$ for polynomials and $\zeta_K = O(K^{1/2})$ for splines.

Assumption 18.7.1

1. *For some $\alpha > 0$ the series approximation satisfies (18.3).*
2. $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) \le \bar{\sigma}^2 < \infty$.
3. $\lambda_{\min}(\boldsymbol{Q}_K) \ge \underline{\lambda} > 0$.
4. *$K = K(n)$ is a function of n which satisfies $K/n \to 0$ and $\zeta_K^2 K/n \to 0$ as $n \to \infty$.*

Assumptions 18.7.1.1 through 18.7.1.3 concern properties of the regression model. Assumption 18.7.1.1 holds with $\alpha = s/d$ if $\mathcal{X}$ is compact and the s'th derivative of $m(\boldsymbol{x})$ is continuous. Assumption 18.7.1.2 allows for conditional heteroskedasticity, but requires the conditional variance to be bounded. Assumption 18.7.1.3 excludes near-singular designs. Since estimates of the conditional mean are unchanged if we replace $\boldsymbol{z}_{Ki}$ with $\boldsymbol{z}_{Ki}^* = \boldsymbol{B}_K\boldsymbol{z}_{Ki}$ for any non-singular $\boldsymbol{B}_K$, Assumption 18.7.1.3 can be viewed as holding after transformation by an appropriate non-singular $\boldsymbol{B}_K$.

Assumption 18.7.1.4 concerns the choice of the number of series terms, which is under the control of the user. It specifies that K can increase with sample size, but at a controlled rate of growth. Since $\zeta_K = O(K)$ for polynomials and $\zeta_K = O(K^{1/2})$ for splines, Assumption 18.7.1.4 is satisfied if $K^3/n \to 0$ for polynomials and $K^2/n \to 0$ for splines. This means that while the number of series terms K can increase with the sample size, K must increase at a much slower rate.

In Section 18.5 we introduced the best uniform approximation, and in this section we introduced the best linear predictor. What is the relationship? They may be similar in practice, but they are not the same and we should be careful to maintain the distinction. Note that from (18.5) we can write $m(\boldsymbol{x}_i) = \boldsymbol{z}_{Ki}'\boldsymbol{\beta}_K^* + r_{Ki}^*$ where $r_{Ki}^* = r_K^*(\boldsymbol{x}_i)$ satisfies $\sup_i |r_{Ki}^*| = O\left(K^{-\alpha}\right)$ from (18.6). Then the best linear predictor equals

$$\begin{aligned}
\boldsymbol{\beta}_K &= \mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}y_i\right) \\
&= \mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}m(\boldsymbol{x}_i)\right) \\
&= \mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}(\boldsymbol{z}_{Ki}'\boldsymbol{\beta}_K^* + r_{Ki}^*)\right) \\
&= \boldsymbol{\beta}_K^* + \mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}r_{Ki}^*\right).
\end{aligned}$$

Thus the difference between the two approximations is

$$\begin{aligned}
r_K(\boldsymbol{x}) - r_K^*(\boldsymbol{x}) &= \boldsymbol{z}_K(\boldsymbol{x})'\left(\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K\right) \\
&= \boldsymbol{z}_K(\boldsymbol{x})'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}r_{Ki}^*\right). \qquad (18.12)
\end{aligned}$$

Observe that by the properties of projection

$$\mathbb{E}\left(\boldsymbol{r}_{Ki}^{*2}\right) - \mathbb{E}\left(\boldsymbol{r}_{Ki}^*\boldsymbol{z}_{Ki}\right)'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}r_{Ki}^*\right) \geq 0 \qquad (18.13)$$

and by (18.6)

$$\mathbb{E}\left(r_{Ki}^{*2}\right) = \int r_K^*(\boldsymbol{x})^2 f_x(\boldsymbol{x})d\boldsymbol{x} \leq O\left(K^{-2\alpha}\right). \qquad (18.14)$$

Then applying the Schwarz inequality to (18.12), Definition (18.11), (18.13) and (18.14), we find

$$\begin{aligned}
|r_K(\boldsymbol{x}) - r_K^*(\boldsymbol{x})| &\leq \left(\boldsymbol{z}_K(\boldsymbol{x})'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\boldsymbol{z}_K(\boldsymbol{x})\right)^{1/2} \\
&\quad \left(\mathbb{E}\left(r_{Ki}^*\boldsymbol{z}_{Ki}\right)'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}r_{Ki}^*\right)\right)^{1/2} \\
&\leq O\left(\zeta_K K^{-\alpha}\right). \qquad (18.15)
\end{aligned}$$

It follows that the best linear predictor approximation error satisfies

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |r_K(\boldsymbol{x})| \leq O\left(\zeta_K K^{-\alpha}\right). \qquad (18.16)$$

The bound (18.16) is probably not the best possible, but it shows that the best linear predictor satisfies a uniform approximation bound. Relative to (18.6), the rate is slower by the factor ζ_K. The bound (18.16) term is $o(1)$ as $K \to \infty$ if $\zeta_K K^{-\alpha} \to 0$. A sufficient condition is that $\alpha > 1$ ($s > d$) for polynomials and $\alpha > 1/2$ ($s > d/2$) for splines, where $d = \dim(\boldsymbol{x})$ and s is the number of continuous derivatives of $m(\boldsymbol{x})$.

It is also useful to observe that since $\boldsymbol{\beta}_K$ is the best linear approximation to $m(\boldsymbol{x}_i)$ in mean-square (see Section 2.24), then

$$\begin{aligned}
\mathbb{E}\left(r_{Ki}^2\right) &= \mathbb{E}\left(\left(m(\boldsymbol{x}_i) - \boldsymbol{z}_{Ki}'\boldsymbol{\beta}_K\right)^2\right) \\
&\leq \mathbb{E}\left(\left(m(\boldsymbol{x}_i) - \boldsymbol{z}_{Ki}'\boldsymbol{\beta}_K^*\right)^2\right) \\
&\leq O\left(K^{-2\alpha}\right) \qquad (18.17)
\end{aligned}$$

the final inequality by (18.14).

18.8 Residuals and Regression Fit

The fitted regression at $\boldsymbol{x} = \boldsymbol{x}_i$ is $\widehat{m}_K(\boldsymbol{x}_i) = \boldsymbol{z}_{Ki}'\widehat{\boldsymbol{\beta}}_K$ and the fitted residual is

$$\widehat{e}_{iK} = y_i - \widehat{m}_K(\boldsymbol{x}_i).$$

The leave-one-out prediction errors are

$$\begin{aligned}\widetilde{e}_{iK} &= y_i - \widehat{m}_{K,-i}(\boldsymbol{x}_i) \\ &= y_i - \boldsymbol{z}_{Ki}'\widehat{\boldsymbol{\beta}}_{K,-i}\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{K,-i}$ is the least-squares coefficient with the i'th observation omitted. Using (3.44) we can also write

$$\widetilde{e}_{iK} = \widehat{e}_{iK}(1 - h_{Kii})^{-1}$$

where $h_{Kii} = \boldsymbol{z}_{Ki}'\left(\boldsymbol{Z}_K'\boldsymbol{Z}_K\right)^{-1}\boldsymbol{z}_{Ki}$.

As for kernel regression, the prediction errors $\widetilde{e}_{iK}$ are better estimates of the errors than the fitted residuals $\widehat{e}_{iK}$, as they do not have the tendency to over-fit when the number of series terms is large.

To assess the fit of the nonparametric regression, the estimate of the mean-square prediction error is

$$\widetilde{\sigma}_K^2 = \frac{1}{n}\sum_{i=1}^n \widetilde{e}_{iK}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{e}_{iK}^2(1 - h_{Kii})^{-2}$$

and the prediction R^2 is

$$\widetilde{R}_K^2 = 1 - \frac{\sum_{i=1}^n \widetilde{e}_{iK}^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

18.9 Cross-Validation Model Selection

The cross-validation criterion for selection of the number of series terms is the MSPE

$$CV(K) = \widetilde{\sigma}_K^2 = \frac{1}{n}\sum_{i=1}^n \widehat{e}_{iK}^2(1 - h_{Kii})^{-2}.$$

By selecting the series terms to minimize $CV(K)$, or equivalently maximize $\widetilde{R}_K^2$, we have a data-dependent rule which is designed to produce estimates with low integrated mean-squared error (IMSE) and mean-squared forecast error (MSFE). As shown in Theorem 17.6.1, $CV(K)$ is an approximately unbiased estimated of the MSFE and IMSE, so finding the model which produces the smallest value of $CV(K)$ is a good indicator that the estimated model has small MSFE and IMSE. The proof of the result is the same for all nonparametric estimators (series as well as kernels) so does not need to be repeated here.

As a practical matter, an estimator corresponds to a set of regressors $\boldsymbol{z}_{Ki}$, that is, a set of transformations of the original variables $\boldsymbol{x}_i$. For each set of regressions, the regression is estimated and $CV(K)$ calculated, and the estimator is selected which has the smallest value of $CV(K)$. If there are p ordered regressors, then there are p possible estimators. Typically, this calculation is simple even if p is large. However, if the p regressors are unordered (and this is typical) then there are 2^p possible subsets of conceivable models. If p is even moderately large, 2^p can be immensely large so brute-force computation of all models may be computationally demanding.

18.10 Convergence in Mean-Square

The series estimate $\widehat{\boldsymbol{\beta}}_K$ are indexed by K. The point of nonparametric estimation is to let K be flexible so as to incorporate greater complexity when the data are sufficiently informative. This means that K will typically be increasing with sample size n. This invalidates conventional asymptotic distribution theory. However, we can develop extensions which use appropriate matrix norms, and by focusing on real-valued functions of the parameters including the estimated regression function itself.

The asymptotic theory we present in this and the next several sections is largely taken from Newey (1997).

Our first main result shows that the least-squares estimate converges to $\boldsymbol{\beta}_K$ in mean-square distance.

Theorem 18.10.1 *Under Assumption 18.7.1, as* $n \to \infty$,

$$\left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)' \boldsymbol{Q}_K \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right) = O_p\left(\frac{K}{n}\right) + o_p\left(K^{-2\alpha}\right) \tag{18.18}$$

The proof of Theorem 18.10.1 is rather technical and deferred to Section 18.16.

The rate of convergence in (18.18) has two terms. The $O_p\left(K/n\right)$ term is due to estimation variance. Note in contrast that the corresponding rate would be $O_p\left(1/n\right)$ in the parametric case. The difference is that in the parametric case we assume that the number of regressors K is fixed as n increases, while in the nonparametric case we allow the number of regressors K to be flexible. As K increases, the estimation variance increases. The $o_p\left(K^{-2\alpha}\right)$ term in (18.18) is due to the series approximation error.

Using Theorem 18.10.1 we can establish the following convergence rate for the estimated regression function.

Theorem 18.10.2 *Under Assumption 18.7.1, as* $n \to \infty$,

$$\int \left(\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})\right)^2 f_x(\boldsymbol{x}) d\boldsymbol{x} = O_p\left(\frac{K}{n}\right) + O_p\left(K^{-2\alpha}\right) \tag{18.19}$$

Theorem 18.10.2 shows that the integrated squared difference between the fitted regression and the true CEF converges in probability to zero if $K \to \infty$ as $n \to \infty$. The convergence results of Theorem 18.10.2 show that the number of series terms K involves a trade-off similar to the role of the bandwidth h in kernel regression. Larger K implies smaller approximation error but increased estimation variance.

The optimal rate which minimizes the average squared error in (18.19) is $K = O\left(n^{1/(1+2\alpha)}\right)$, yielding an optimal rate of convergence in (18.19) of $O_p\left(n^{-2\alpha/(1+2\alpha)}\right)$. This rate depends on the unknown smoothness α of the true CEF (the number of derivatives s) and so does not directly syggest a practical rule for determining K. Still, the implication is that when the function being estimated is less smooth (α is small) then it is necessary to use a larger number of series terms K to reduce the bias. In contrast, when the function is more smooth then it is better to use a smaller number of series terms K to reduce the variance.

To establish (18.19), using (18.7) and (18.8) we can write

$$\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})' \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right) - r_K(\boldsymbol{x}). \tag{18.20}$$

Since e_{Ki} are projection errors, they satisfy $\mathbb{E}(\boldsymbol{z}_{Ki}e_{Ki}) = 0$ and thus $\mathbb{E}(\boldsymbol{z}_{Ki}r_{Ki}) = 0$. This means $\int \boldsymbol{z}_K(\boldsymbol{x})r_K(\boldsymbol{x})f_x(\boldsymbol{x})d\boldsymbol{x} = 0$. Also observe that $\boldsymbol{Q}_K = \int \boldsymbol{z}_K(\boldsymbol{x})\boldsymbol{z}_K(\boldsymbol{x})'f_x(\boldsymbol{x})d\boldsymbol{x}$ and $\mathbb{E}\left(r_{Ki}^2\right) = \int r_K(\boldsymbol{x})^2 f_x(\boldsymbol{x})d\boldsymbol{x}$. Then

$$\begin{aligned}
&\int \left(\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})\right)^2 f_x(\boldsymbol{x})d\boldsymbol{x} \\
&= \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)' \boldsymbol{Q}_K \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right) + \mathbb{E}\left(r_{Ki}^2\right) \\
&\leq O_p\left(\frac{K}{n}\right) + O_p\left(K^{-2\alpha}\right)
\end{aligned}$$

by (18.18) and (18.17), establishing (18.19).

18.11 Uniform Convergence

Theorem 18.10.2 established conditions under which $\widehat{m}_K(\boldsymbol{x})$ is consistent in a squared error norm. It is also of interest to know the rate at which the largest deviation converges to zero. We have the following rate.

Theorem 18.11.1 *Under Assumption 18.7.1, then as $n \to \infty$,*

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})| = O_p\left(\sqrt{\frac{\zeta_K^2 K}{n}}\right) + O_p\left(\zeta_K K^{-\alpha}\right). \tag{18.21}$$

Relative to Theorem 18.10.2, the error has been increased multiplicatively by ζ_K. This slower convergence rate is a penalty for the stronger uniform convergence, though it is probably not the best possible rate. Examining the bound in (18.21) notice that the first term is $o_p(1)$ under Assumption 18.7.1.4. The second term is $o_p(1)$ if $\zeta_K K^{-\alpha} \to 0$, which requires that $K \to \infty$ and that α be sufficiently large. A sufficient condition is that $s > d$ for polynomials and $s > d/2$ for splines, where $d = \dim(\boldsymbol{x})$ and s is the number of continuous derivatives of $m(\boldsymbol{x})$. Thus higher dimensional $\boldsymbol{x}$ require a smoother CEF $m(\boldsymbol{x})$ to ensure that the series estimate $\widehat{m}_K(\boldsymbol{x})$ is uniformly consistent.

The convergence (18.21) is straightforward to show using (18.18). Using (18.20), the Triangle Inequality, the Schwarz inequality (A.20), Definition (18.11), (18.18) and (18.16),

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathcal{X}} |\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\boldsymbol{z}_K(\boldsymbol{x})'\left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)\right| + \sup_{\boldsymbol{x}\in\mathcal{X}} |r_K(\boldsymbol{x})| \\
&\leq \sup_{\boldsymbol{x}\in\mathcal{X}} \left(\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{Q}_K^{-1}\boldsymbol{z}_K(\boldsymbol{x})\right)^{1/2} \left(\left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)' \boldsymbol{Q}_K \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)\right)^{1/2} \\
&+ O\left(\zeta_K K^{-\alpha}\right) \\
&\leq \zeta_K \left(O_p\left(\frac{K}{n}\right) + O_p\left(K^{-2\alpha}\right)\right)^{1/2} + O\left(\zeta_K K^{-\alpha}\right), \\
&= O_p\left(\sqrt{\frac{\zeta_K^2 K}{n}}\right) + O_p\left(\zeta_K K^{-\alpha}\right).
\end{aligned} \tag{18.22}$$

This is (18.21).

18.12 Asymptotic Normality

One advantage of series methods is that the estimators are (in finite samples) equivalent to parametric estimators, so it is easy to calculate covariance matrix estimates. We now show that we can also justify normal asymptotic approximations.

The theory we present in this section will apply to any linear function of the regression function. That is, we allow the parameter of interest to be aany non-trivial real-valued linear function of the entire regression function $m(\cdot)$

$$\theta = a(m).$$

This includes the regression function $m(\boldsymbol{x})$ at a given point $\boldsymbol{x}$, derivatives of $m(\boldsymbol{x})$, and integrals over $m(\boldsymbol{x})$. Given $\widehat{m}_K(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})'\widehat{\boldsymbol{\beta}}_K$ as an estimator for $m(\boldsymbol{x})$, the estimator for θ is

$$\widehat{\theta}_K = a(\widehat{m}_K) = \boldsymbol{a}_K'\widehat{\boldsymbol{\beta}}_K$$

for some $K \times 1$ vector of constants $\boldsymbol{a}_K \neq \boldsymbol{0}$. (The relationship $a(\widehat{m}_K) = \boldsymbol{a}_K'\widehat{\boldsymbol{\beta}}_K$ follows since a is linear in m and $\widehat{m}_K$ is linear in $\widehat{\boldsymbol{\beta}}_K$.)

If K were fixed as $n \to \infty$, then by standard asymptotic theory we would expect $\widehat{\theta}_K$ to be asymptotically normal with variance

$$v_K = \boldsymbol{a}_K'\boldsymbol{Q}_K^{-1}\boldsymbol{\Omega}_K\boldsymbol{Q}_K^{-1}\boldsymbol{a}_K$$

where

$$\boldsymbol{\Omega}_K = \mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'e_{Ki}^2\right).$$

The standard justification, however, is not valid in the nonparametric case, in part because v_K may diverge as $K \to \infty$, and in part due to the finite sample bias due to the approximation error. Therefore a new theory is required. Interestingly, it turns out that in the nonparametric case $\widehat{\theta}_K$ is still asymptotically normal, and v_K is still the appropriate variance for $\widehat{\theta}_K$. The proof is different than the parametric case as the dimensions of the matrices are increasing with K, and we need to be attentive to the estimator's bias due to the series approximation.

Theorem 18.12.1 *Under Assumption 18.7.1, if in addition* $\mathbb{E}\left(e_i^4|\boldsymbol{x}_i\right) \leq \kappa_4 < \infty$, $\mathbb{E}\left(e_i^2|\boldsymbol{x}_i\right) \geq \underline{\sigma}^2 > 0$, *and* $\zeta_K K^{-\alpha} = O(1)$, *then as* $n \to \infty$,

$$\frac{\sqrt{n}\left(\widehat{\theta}_K - \theta + a(r_K)\right)}{v_K^{1/2}} \xrightarrow{d} \mathrm{N}(0,1) \tag{18.23}$$

The proof of Theorem 18.12.1 can be found in Section 18.16.

Theorem 18.12.1 shows that the estimator $\widehat{\theta}_K$ is approximately normal with bias $-a(r_K)$ and variance v_K/n. The variance is the same as in the parametric case, but the asymptotic distribution contains an asymptotic bias, similar as is found in kernel regression. We discuss the bias in more detail below.

Notice that Theorem 18.12.1 requires $\zeta_K K^{-\alpha} = O(1)$, which is similar to that found in Theorem 18.11.1 to establish uniform convergence. The the bound $\zeta_K K^{-\alpha} = O(1)$ allows K to be constant with n or to increase with n. However, when K is increasing the bound requires that α be sufficient large so that K^α grows faster than ζ_K. A sufficient condition is that $s = d$ for polynomials and $s = d/2$ for splines. The fact that the condition allows for K to be constant means that Theorem 18.12.1 includes parametric least-squares as a special case with explicit attention to estimation bias.

One useful message from Theorem 18.12.1 is that the classic variance formula v_K for $\widehat{\theta}_K$ still applies for series regression. Indeed, we can estimate the asymptotic variance using the standard White formula

$$\widehat{\nu}_K = \boldsymbol{a}_K' \widehat{\boldsymbol{Q}}_K^{-1} \widehat{\boldsymbol{\Omega}}_K \widehat{\boldsymbol{Q}}_K^{-1} \boldsymbol{a}_K$$
$$\widehat{\boldsymbol{\Omega}}_K = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_{Ki} \boldsymbol{z}_{Ki}' \widehat{e}_{iK}^2$$
$$\widehat{\boldsymbol{Q}}_K = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{z}_{Ki} \boldsymbol{z}_{Ki}'.$$

Hence a standard error for $\hat{\theta}_K$ is

$$\hat{s}(\theta_K) = \sqrt{\frac{1}{n} \boldsymbol{a}_K' \widehat{\boldsymbol{Q}}_K^{-1} \widehat{\boldsymbol{\Omega}}_K \widehat{\boldsymbol{Q}}_K^{-1} \boldsymbol{a}_K}.$$

It can be shown (Newey, 1997) that $\widehat{\nu}_K / v_K \xrightarrow{p} 1$ as $n \to \infty$ and thus the distribution in (18.23) is unchanged if v_K is replaced with $\hat{v}_K$.

Theorem 18.12.1 shows that the estimator $\widehat{\theta}_K$ has a bias term $a\left(r_K\right)$. What is this? It is the same transformation of the function $r_K(\boldsymbol{x})$ as $\theta = a\left(m\right)$ is of the regression function $m(\boldsymbol{x})$. For example, if $\theta = m(\boldsymbol{x})$ is the regression at a fixed point $\boldsymbol{x}$, then $a\left(r_K\right) = r_K(\boldsymbol{x})$, the approximation error at the same point. If $\theta = \dfrac{d}{dx} m(x)$ is the regression derivative, then $a\left(r_K\right) = \dfrac{d}{dx} r_K(\boldsymbol{x})$ is the derivative of the approximation error.

This means that the bias in the estimator $\widehat{\theta}_K$ for θ shown in Theorem 18.12.1 is simply the approximation error, transformed by the functional of interest. If we are estimating the regression function then the bias is the error in approximating the regression function; if we are estimating the regression derivative then the bias is the error in the derivative in the approximation error for the regression function.

18.13 Asymptotic Normality with Undersmoothing

An unpleasant aspect about Theorem 18.12.1 is the bias term. An interesting trick is that this bias term can be made asymptotically negligible if we assume that K increases with n at a sufficiently fast rate.

Theorem 18.13.1 *Under Assumption 18.7.1, if in addition* $\mathbb{E}\left(e_i^4 | \boldsymbol{x}_i\right) \leq \kappa_4 < \infty$, $\mathbb{E}\left(e_i^2 | \boldsymbol{x}_i\right) \geq \underline{\sigma}^2 > 0$, $a\left(r_K^*\right) \leq O\left(K^{-\alpha}\right)$, $nK^{-2\alpha} \to 0$, *and* $\boldsymbol{a}_K' \boldsymbol{Q}_K^{-1} \boldsymbol{a}_K$ *is bounded away from zero, then*

$$\frac{\sqrt{n}\left(\widehat{\theta}_K - \theta\right)}{v_K^{1/2}} \xrightarrow{d} \mathrm{N}\left(0, 1\right). \tag{18.24}$$

The condition $a\left(r_K^*\right) \leq O\left(K^{-\alpha}\right)$ states that the function of interest (for example, the regression function, its derivative, or its integral) applied to the uniform approximation error converges to zero as the number of terms K in the series approximation increases. If $a\left(m\right) = m(\boldsymbol{x})$ then this condition holds by (18.6).

The condition that $\boldsymbol{a}_K' \boldsymbol{Q}_K^{-1} \boldsymbol{a}_K$ is bounded away from zero is simply a technical requirement to exclude degeneracy.

The critical condition is the assumption that $nK^{-2\alpha} \to 0$. This requires that $K \to \infty$ at a rate *faster* than $n^{1/2\alpha}$. This is a troubling condition. The optimal rate for estimation of $m(\boldsymbol{x})$ is $K = O\left(n^{1/(1+2\alpha)}\right)$. If we set $K = n^{1/(1+2\alpha)}$ by this rule then $nK^{-2\alpha} = n^{1/(1+2\alpha)} \to \infty$, not zero. Thus this assumption is equivalent to assuming that K is much larger than optimal. The reason why this trick works (that is, why the bias is negligible) is that by increasing K, the asymptotic bias decreases and the asymptotic variance increases and thus the variance dominates. Because K is larger than optimal, we typically say that $\widehat{m}_K(\boldsymbol{x})$ is **undersmoothed** relative to the optimal series estimator.

Many authors like to focus their asymptotic theory on the assumptions in Theorem 18.13.1, as the distribution (18.24) appears cleaner. However, it is a poor use of asymptotic theory. There are three problems with the assumption $nK^{-2\alpha} \to 0$ and the approximation (18.24). First, it says that if we intentionally pick K to be larger than optimal, we can increase the estimation variance relative to the bias so the variance will dominate the bias. But why would we want to intentionally use an estimator which is sub-optimal? Second, the assumption $nK^{-2\alpha} \to 0$ does not eliminate the asymptotic bias, it only makes it of lower order than the variance. So the approximation (18.24) is technically valid, but the missing asymptotic bias term is just slightly smaller in asymptotic order, and thus still relevant in finite samples. Third, the condition $nK^{-2\alpha} \to 0$ is just an assumption, it has nothing to do with actual empirical practice. Thus the difference between (18.23) and (18.24) is in the assumptions, not in the actual reality or in the actual empirical practice. Eliminating a nuisance (the asymptotic bias) through an assumption is a trick, not a substantive use of theory. My strong view is that the result (18.23) is more informative than (18.24). It shows that the asymptotic distribution is normal but has a non-trivial finite sample bias.

18.14 Regression Estimation

A special yet important example of a linear estimator of the regression function is the regression function at a fixed point $\boldsymbol{x}$. In the notation of the previous section, $a(m) = m(\boldsymbol{x})$ and $\boldsymbol{a}_K = \boldsymbol{z}_K(\boldsymbol{x})$. The series estimator of $m(\boldsymbol{x})$ is $\hat{\theta}_K = \widehat{m}_K(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})'\widehat{\boldsymbol{\beta}}_K$. As this is a key problem of interest, we restate the asymptotic results of Theorems 18.12.1 and 18.13.1 for this estimator.

Theorem 18.14.1 *Under Assumption 18.7.1, if in addition* $\mathbb{E}\left(e_i^4|\boldsymbol{x}_i\right) \leq \kappa_4 < \infty$, $\mathbb{E}\left(e_i^2|\boldsymbol{x}_i\right) \geq \underline{\sigma}^2 > 0$, *and* $\zeta_K K^{-\alpha} = O(1)$, *then as* $n \to \infty$,

$$\frac{\sqrt{n}\left(\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x}) + \boldsymbol{r}_K(\boldsymbol{x})\right)}{v_K^{1/2}(\boldsymbol{x})} \xrightarrow{d} \mathrm{N}(0,1) \tag{18.25}$$

where

$$v_K(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{Q}_K^{-1}\boldsymbol{\Omega}_K\boldsymbol{Q}_K^{-1}\boldsymbol{z}_K(\boldsymbol{x}).$$

If $\zeta_K K^{-\alpha} = O(1)$ *is replaced by* $nK^{-2\alpha} \to 0$, *and* $\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{Q}_K^{-1}\boldsymbol{z}_K(\boldsymbol{x})$ *is bounded away from zero, then*

$$\frac{\sqrt{n}\left(\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})\right)}{v_K^{1/2}(\boldsymbol{x})} \xrightarrow{d} \mathrm{N}(0,1) \tag{18.26}$$

There are two important features about the asymptotic distribution (18.25).

First, as mentioned in the previous section, it shows how to construct asymptotic standard errors for the CEF $m(\boldsymbol{x})$. These are

$$\hat{s}(\boldsymbol{x}) = \sqrt{\frac{1}{n}\boldsymbol{z}_K(\boldsymbol{x})'\widehat{\boldsymbol{Q}}_K^{-1}\widehat{\boldsymbol{\Omega}}_K\widehat{\boldsymbol{Q}}_K^{-1}\boldsymbol{z}_K(\boldsymbol{x})}.$$

Second, (18.25) shows that the estimator has the asymptotic bias component $\boldsymbol{r}_K(\boldsymbol{x})$. This is due to the fact that the finite order series is an approximation to the unknown CEF $m(\boldsymbol{x})$, and this results in finite sample bias.

The asymptotic distribution (18.26) shows that the bias term is negligable if K diverges fast enough so that $nK^{-2\alpha} \to 0$. As discussed in the previous section, this means that K is larger than optimal.

The assumption that $\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{Q}_K^{-1}\boldsymbol{z}_K(\boldsymbol{x})$ is bounded away from zero is a technical condition to exclude degenerate cases, and is automatically satisfied if $\boldsymbol{z}_K(\boldsymbol{x})$ includes an intercept.

Plots of the CEF estimate $\widehat{m}_K(\boldsymbol{x})$ can be accompanied by 95% confidence intervals $\widehat{m}_K(\boldsymbol{x}) \pm 2\hat{s}(\boldsymbol{x})$. As we discussed in the chapter on kernel regression, this can be viewed as a confidence interval for the pseudo-true CEF $m_K^*(\boldsymbol{x}) = m(\boldsymbol{x}) - \boldsymbol{r}_K(\boldsymbol{x})$, not for the true $m(\boldsymbol{x})$. As for kernel regression, the difference is the unavoidable consequence of nonparametric estimation.

18.15 Kernel Versus Series Regression

In this and the previous chapter we have presented two distinct methods of nonparametric regression based on kernel methods and series methods. Which should be used in practice? Both methods have advantages and disadvantages and there is no clear overall winner.

First, while the asymptotic theory of the two estimators appear quite different, they are actually rather closely related. When the regression function $m(\boldsymbol{x})$ is twice differentiable ($s = 2$) then the rate of convergence of both the MSE of the kernel regression estimator with optimal bandwidth h and the series estimator with optimal K is $n^{-2/(d+4)}$. There is no difference. If the regression function is smoother than twice differentiable ($s > 2$) then the rate of the convergence of the series estimator improves. This may appear to be an advantage for series methods, but kernel regression can also take advantage of the higher smoothness by using so-called higher-order kernels or local polynomial regression, so perhaps this advantage is not too large.

Both estimators are asymptotically normal and have straightforward asymptotic standard error formulae. The series estimators are a bit more convenient for this purpose, as classic parametric standard error formula work without amendment.

An advantage of kernel methods is that their distributional theory is easier to derive. The theory is all based on local averages which is relatively straightforward. In contrast, series theory is more challenging, dealing with increasing parameter spaces. An important difference in the theory is that for kernel estimators we have explicit representations for the bias while we only have rates for series methods. This means that plug-in methods can be used for bandwidth selection in kernel regression. However, typically we rely on cross-validation, which is equally applicable in both kernel and series regression.

Kernel methods are also relatively easy to implement when the dimension d is large. There is not a major change in the methodology as d increases. In contrast, series methods become quite cumbersome as d increases as the number of cross-terms increases exponentially.

A major advantage of series methods is that it has inherently a high degree of flexibility, and the user is able to implement shape restrictions quite easily. For example, in series estimation it is relatively simple to implement a partial linear CEF, an additively separable CEF, monotonicity, concavity or convexity. These restrictions are harder to implement in kernel regression.

18.16 Technical Proofs

Define $\boldsymbol{z}_{Ki} = \boldsymbol{z}_K(\boldsymbol{x}_i)$ and let $\boldsymbol{Q}_K^{1/2}$ denote the positive definite square root of $\boldsymbol{Q}_K$. As mentioned before Theorem 18.10.1, the regression problem is unchanged if we replace $\boldsymbol{z}_{Ki}$ with a rotated regressor such as $\boldsymbol{z}_{Ki}^* = \boldsymbol{Q}_K^{-1/2}\boldsymbol{z}_{Ki}$. This is a convenient choice for then $\mathbb{E}\left(\boldsymbol{z}_{Ki}^*\boldsymbol{z}_{Ki}^{*\prime}\right) = \boldsymbol{I}_K$. For notational convenience we will simply write the transformed regressors as $\boldsymbol{z}_{Ki}$ and set $\boldsymbol{Q}_K = \boldsymbol{I}_K$.

We start with some convergence results for the sample design matrix

$$\widehat{\boldsymbol{Q}}_K = \frac{1}{n}\boldsymbol{Z}'_K\boldsymbol{Z}_K = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_{Ki}\boldsymbol{z}'_{Ki}.$$

Theorem 18.16.1 *Under Assumption 18.7.1 and $\boldsymbol{Q}_K = \boldsymbol{I}_K$, as $n \to \infty$,*

$$\left\|\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K\right\| = o_p(1) \tag{18.27}$$

and

$$\lambda_{\min}(\widehat{\boldsymbol{Q}}_K) \xrightarrow{p} 1. \tag{18.28}$$

Proof. Since

$$\left\|\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K\right\|^2 = \sum_{j=1}^{K}\sum_{\ell=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{z}_{jKi}\boldsymbol{z}_{\ell Ki} - \mathbb{E}\left(\boldsymbol{z}_{jKi}\boldsymbol{z}_{\ell Ki}\right)\right)\right)^2$$

then

$$\begin{aligned}
\mathbb{E}\left(\left\|\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K\right\|^2\right) &= \sum_{j=1}^{K}\sum_{\ell=1}^{K}\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_{jKi}\boldsymbol{z}_{\ell Ki}\right) \\
&= n^{-1}\sum_{j=1}^{K}\sum_{\ell=1}^{K}\operatorname{var}\left(\boldsymbol{z}_{jKi}\boldsymbol{z}_{\ell Ki}\right) \\
&\leq n^{-1}\mathbb{E}\left(\sum_{j=1}^{K}\boldsymbol{z}_{jKi}^2\sum_{\ell=1}^{K}\boldsymbol{z}_{\ell Ki}^2\right) \\
&= n^{-1}\mathbb{E}\left(\left(\boldsymbol{z}'_{Ki}\boldsymbol{z}_{Ki}\right)^2\right).
\end{aligned} \tag{18.29}$$

Since $\boldsymbol{z}'_{Ki}\boldsymbol{z}_{Ki} \leq \zeta_K^2$ by definition (18.11) and using (A.1) we find

$$\mathbb{E}\left(\boldsymbol{z}'_{Ki}\boldsymbol{z}_{Ki}\right) = \operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}'_{Ki}\right)\right) = \operatorname{tr}\boldsymbol{I}_K = K, \tag{18.30}$$

so that

$$\mathbb{E}\left(\left(\boldsymbol{z}'_{Ki}\boldsymbol{z}_{Ki}\right)^2\right) \leq \zeta_K^2 K \tag{18.31}$$

and hence (18.29) is $o(1)$ under Assumption 18.7.1.4. Theorem 6.13.1 shows that this implies (18.27).

Let $\lambda_1, \lambda_2, ..., \lambda_K$ be the eigenvalues of $\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K$ which are real as $\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K$ is symmetric. Then

$$\left|\lambda_{\min}(\widehat{\boldsymbol{Q}}_K) - 1\right| = \left|\lambda_{\min}(\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K)\right| \leq \left(\sum_{\ell=1}^{K}\lambda_\ell^2\right)^{1/2} = \left\|\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K\right\|$$

where the second equality is (A.22). This is $o_p(1)$ by (18.27), establishing (18.28). ■

Proof of Theorem 18.10.1. As above, assume that the regressors have been transformed so that $\boldsymbol{Q}_K = \boldsymbol{I}_K$.

From expression (18.10) we can substitute to find

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K &= \left(\boldsymbol{Z}'_K \boldsymbol{Z}_K\right)^{-1} \boldsymbol{Z}'_K \boldsymbol{e}_K. \\ &= \widehat{\boldsymbol{Q}}_K^{-1} \left(\frac{1}{n} \boldsymbol{Z}'_K \boldsymbol{e}_K\right) \end{aligned} \tag{18.32}$$

Using (18.32) and the Quadratic Inequality (A.28),

$$\begin{aligned}&\left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)' \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right) \\ &= n^{-2} \left(\boldsymbol{e}'_K \boldsymbol{Z}_K\right) \widehat{\boldsymbol{Q}}_K^{-1} \widehat{\boldsymbol{Q}}_K^{-1} \left(\boldsymbol{Z}'_K \boldsymbol{e}_K\right) \\ &\leq \left(\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}\right)\right)^2 n^{-2} \left(\boldsymbol{e}'_K \boldsymbol{Z}_K \boldsymbol{Z}'_K \boldsymbol{e}_K\right). \end{aligned} \tag{18.33}$$

Observe that (18.28) implies

$$\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}\right) = \left(\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K\right)\right)^{-1} = O_p(1). \tag{18.34}$$

Since $e_{Ki} = e_i + r_{Ki}$, and using Assumption 18.7.1.2 and (18.16), then

$$\sup_i \mathbb{E}\left(e_{Ki}^2 | \boldsymbol{x}_i\right) = \overline{\sigma}^2 + \sup_i r_{Ki}^2 \leq \overline{\sigma}^2 + O\left(\zeta_K^2 K^{-2\alpha}\right). \tag{18.35}$$

As e_{Ki} are projection errors, they satisfy $\mathbb{E}\left(\boldsymbol{z}_{Ki} e_{Ki}\right) = 0$. Since the observations are independent, using (18.30) and (18.35), then

$$\begin{aligned} n^{-2}\mathbb{E}\left(\boldsymbol{e}'_K \boldsymbol{Z}_K \boldsymbol{Z}'_K \boldsymbol{e}_K\right) &= n^{-2}\mathbb{E}\left(\sum_{i=1}^n e_{Ki} \boldsymbol{z}'_{Ki} \sum_{ij=1}^n \boldsymbol{z}_{Kj} e_{Kj}\right) \\ &= n^{-2} \sum_{i=1}^n \mathbb{E}\left(\boldsymbol{z}'_{Ki} \boldsymbol{z}_{Ki} e_{Ki}^2\right) \\ &\leq n^{-1}\mathbb{E}\left(\boldsymbol{z}'_{Ki} \boldsymbol{z}_{Ki}\right) \sup_i \mathbb{E}\left(e_{Ki}^2 | \boldsymbol{x}_i\right) \\ &\leq \overline{\sigma}^2 \frac{K}{n} + O\left(\frac{\zeta_K^2 K^{1-2\alpha}}{n}\right) \\ &= \overline{\sigma}^2 \frac{K}{n} + o\left(K^{-2\alpha}\right) \end{aligned} \tag{18.36}$$

since $\zeta_K^2 K/n = o(1)$ by Assumption 18.7.1.4. Theorem 6.13.1 shows that this implies

$$n^{-2} \boldsymbol{e}'_K \boldsymbol{Z}_K \boldsymbol{Z}'_K \boldsymbol{e}_K = O_p\left(n^{-2}\right) + o_p\left(K^{-2\alpha}\right). \tag{18.37}$$

Together, (18.33), (18.34) and (18.37) imply (18.18). ■

Proof of Theorem 18.12.1. As above, assume that the regressors have been transformed so that $\boldsymbol{Q}_K = \boldsymbol{I}_K$.

Using $m(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K + r_K(\boldsymbol{x})$ and linearity

$$\begin{aligned}\theta &= a\left(m\right) \\ &= a\left(\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K\right) + a\left(r_K\right) \\ &= \boldsymbol{a}'_K \boldsymbol{\beta}_K + a\left(r_K\right) \end{aligned}$$

Combined with (18.32) we find

$$\begin{aligned}\widehat{\theta}_K - \theta + a\left(r_K\right) &= \boldsymbol{a}_K'\left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right) \\ &= \frac{1}{n}\boldsymbol{a}_K'\widehat{\boldsymbol{Q}}_K^{-1}\boldsymbol{Z}_K'\boldsymbol{e}_K\end{aligned}$$

and thus

$$\begin{aligned}\sqrt{\frac{n}{v_k}}\left(\widehat{\theta}_K - \theta_K + a\left(r_K\right)\right) &= \sqrt{\frac{n}{v_k}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right) \\ &= \sqrt{\frac{1}{nv_k}}\boldsymbol{a}_K'\widehat{\boldsymbol{Q}}_K^{-1}\boldsymbol{Z}_K'\boldsymbol{e}_K \\ &= \frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\boldsymbol{Z}_K'\boldsymbol{e}_K \qquad (18.38) \\ &+ \frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1} - \boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{e} \qquad (18.39) \\ &+ \frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1} - \boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{r}_K. \qquad (18.40)\end{aligned}$$

where we have used $\boldsymbol{e}_K = \boldsymbol{e} + \boldsymbol{r}_K$. We now take the terms in (18.38)-(18.40) separately.

First, take (18.38). We can write

$$\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\boldsymbol{Z}_K'\boldsymbol{e}_K = \frac{1}{\sqrt{nv_K}}\sum_{i=1}^{n}\boldsymbol{a}_K'\boldsymbol{z}_{Ki}e_{Ki}. \qquad (18.41)$$

Observe that $\boldsymbol{a}_K'\boldsymbol{z}_{Ki}e_{Ki}$ are independent across i, mean zero, and have variance

$$\mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}e_{Ki}\right)^2\right) = \boldsymbol{a}_K'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'e_{Ki}^2\right)\boldsymbol{a}_K = v_K.$$

We will apply the Lindeberg CLT 6.8.2, for which it is sufficient to verify Lyapunov's condition (6.6):

$$\frac{1}{n^2v_K^2}\sum_{i=1}^{n}\mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}e_{Ki}\right)^4\right) = \frac{1}{nv_K^2}\mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4 e_{Ki}^4\right) \to 0. \qquad (18.42)$$

The assumption that $\zeta_K K^{-\alpha} = O(1)$ means $\zeta_K K^{-\alpha} \le \kappa_1$ for some $\kappa_1 < \infty$. Then by the c_r inequality and $\mathbb{E}\left(e_i^4|\boldsymbol{x}_i\right) \le \kappa$

$$\sup_i \mathbb{E}\left(e_{Ki}^4|\boldsymbol{x}_i\right) \le 8\sup_i\left(\mathbb{E}\left(e_i^4|\boldsymbol{x}_i\right) + r_{Ki}^4\right) \le 8\left(\kappa + \kappa_1\right). \qquad (18.43)$$

Using (18.43), the Schwarz Inequality, and (18.31)

$$\begin{aligned}\mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4 e_{Ki}^4\right) &= \mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4\mathbb{E}\left(e_{Ki}^4|\boldsymbol{x}_i\right)\right) \\ &\le 8\left(\kappa + \kappa_1\right)\mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4\right) \\ &\le 8\left(\kappa + \kappa_1\right)\left(\boldsymbol{a}_K'\boldsymbol{a}_K\right)^2\mathbb{E}\left(\left(\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki}\right)^2\right) \\ &= 8\left(\kappa + \kappa_1\right)\left(\boldsymbol{a}_K'\boldsymbol{a}_K\right)^2\zeta_K^2 K. \qquad (18.44)\end{aligned}$$

Since $\mathbb{E}\left(e_{Ki}^2|\boldsymbol{x}_i\right) = \mathbb{E}\left(e_i^2|\boldsymbol{x}_i\right) + r_{Ki}^2 \ge \underline{\sigma}^2$,

$$\begin{aligned}v_K &= \boldsymbol{a}_K'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'e_{Ki}^2\right)\boldsymbol{a}_K \\ &\ge \underline{\sigma}^2\boldsymbol{a}_K'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)\boldsymbol{a}_K \\ &= \underline{\sigma}^2\boldsymbol{a}_K'\boldsymbol{a}_K. \qquad (18.45)\end{aligned}$$

Equation (18.44) and (18.45) combine to show that

$$\frac{1}{nv_K^2}\mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4 e_{Ki}^4\right) \leq \frac{8\left(\kappa+\kappa_1\right)}{\underline{\sigma}^4}\frac{\zeta_K^2 K}{n} = o(1)$$

under Assumption 18.7.1.4. This establishes Lyapunov's condition (18.42). Hence the Lindeberg CLT applies to (18.41) and we conclude

$$\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\boldsymbol{Z}_K'\boldsymbol{e}_K \xrightarrow{d} \mathrm{N}\left(0,1\right). \tag{18.46}$$

Second, take (18.39). Since $\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = 0$, then applying $\mathbb{E}\left(e_i^2|\boldsymbol{x}_i\right) \leq \bar{\sigma}^2$, the Schwarz and Norm Inequalities, (18.45), (18.34) and (18.27),

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{e}\right)^2 \mid \boldsymbol{X}\right) \\
&= \frac{1}{nv_K}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\boldsymbol{Z}_K\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{a}_K \\
&\leq \frac{\bar{\sigma}^2}{v_K}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\widehat{\boldsymbol{Q}}_K\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{a}_K \\
&= \frac{\bar{\sigma}^2}{v_K}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K-\boldsymbol{I}_K\right)\widehat{\boldsymbol{Q}}_K^{-1}\left(\widehat{\boldsymbol{Q}}_K-\boldsymbol{I}_K\right)\boldsymbol{a}_K \\
&\leq \frac{\bar{\sigma}^2\boldsymbol{a}_K'\boldsymbol{a}_K}{v_K}\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}\right)\left\|\widehat{\boldsymbol{Q}}_K-\boldsymbol{I}_K\right\|^2 \\
&\leq \frac{\bar{\sigma}^2}{\underline{\sigma}^2}o_p(1).
\end{aligned}$$

This establishes

$$\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{e} \xrightarrow{p} 0. \tag{18.47}$$

Third, take (18.40). By the Cauchy-Schwarz inequality, (18.45), and the Quadratic Inequality,

$$\begin{aligned}
&\left(\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{r}_K\right)^2 \\
&\leq \frac{\boldsymbol{a}_K'\boldsymbol{a}_K}{nv_K}\boldsymbol{r}_K'\boldsymbol{Z}_K\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{r}_K \\
&\leq \frac{\boldsymbol{1}}{\underline{\sigma}^2}\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)^2\frac{1}{n}\boldsymbol{r}_K'\boldsymbol{Z}_K\boldsymbol{Z}_K'\boldsymbol{r}_K.
\end{aligned} \tag{18.48}$$

Observe that since the observations are independent and $\mathbb{E}\boldsymbol{z}_{Ki}r_{Ki} = 0$, $\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki} \leq \zeta_K^2$, and (18.17)

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n}\boldsymbol{r}_K'\boldsymbol{Z}_K\boldsymbol{Z}_K'\boldsymbol{r}_K\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n r_{Ki}\boldsymbol{z}_{Ki}'\sum_{ij=1}^n \boldsymbol{z}_{Kj}r_{Kj}\right) \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki}r_{Ki}^2\right) \\
&\leq \zeta_K^2\mathbb{E}\left(r_{Ki}^2\right) \\
&= O\left(\zeta_K^2K^{-2\alpha}\right) \\
&= O(1)
\end{aligned}$$

since $\zeta_K K^{-2} = O(1)$. Thus $\frac{1}{n}\boldsymbol{r}'_K \boldsymbol{Z}_K \boldsymbol{Z}'_K \boldsymbol{r}_K = O_p(1)$. This means that (18.48) is $o_p(1)$ since (18.28) implies

$$\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1} - \boldsymbol{I}_K\right) = \lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}\right) - 1 = o_p(1). \tag{18.49}$$

Equivalently,

$$\frac{1}{\sqrt{n v_K}} \boldsymbol{a}'_K \left(\widehat{\boldsymbol{Q}}_K^{-1} - \boldsymbol{I}_K\right) \boldsymbol{Z}'_K \boldsymbol{r}_K \xrightarrow{p} 0. \tag{18.50}$$

Equations (18.46), (18.47) and (18.50) applied to (18.38)-(18.40) show that

$$\sqrt{\frac{n}{v_k}}\left(\widehat{\theta}_K - \theta_K + a\left(r_K\right)\right) \xrightarrow{d} \mathrm{N}\left(0, 1\right)$$

completing the proof. ■

Proof of Theorem 18.13.1. The assumption that $nK^{-2\alpha} = o(1)$ implies $K^{-\alpha} = o\left(n^{-1/2}\right)$. Thus

$$\zeta_K K^{-\alpha} \leq o\left(\left(\frac{\zeta_K^2}{n}\right)^{1/2}\right) \leq o\left(\left(\frac{\zeta_K^2 K}{n}\right)^{1/2}\right) = o(1)$$

so the conditions of Theorem 18.12.1 are satisfied. It is thus sufficient to show that

$$\sqrt{\frac{n}{v_k}} a\left(r_K\right) = o(1).$$

From (18.12)

$$\begin{aligned} r_K(\boldsymbol{x}) &= r_K^*(\boldsymbol{x}) + \boldsymbol{z}_K(\boldsymbol{x})'\gamma_K \\ \gamma_K &= \mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}'_{Ki}\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki} r^*_{Ki}\right). \end{aligned}$$

Thus by linearity, applying (18.45), and the Schwarz inequality

$$\begin{aligned} \sqrt{\frac{n}{v_k}} a\left(r_K\right) &= \sqrt{\frac{n}{v_k}}\left(a\left(r_K^*\right) + \boldsymbol{a}'_K\gamma_K\right) \\ &\leq \frac{n^{1/2}}{\underline{\sigma}^2\left(\boldsymbol{a}'_K\boldsymbol{a}_K\right)^{1/2}} a\left(r_K^*\right) \qquad (18.51) \\ &\quad + \frac{\left(n\gamma'_K\gamma_K\right)^{1/2}}{\underline{\sigma}}. \qquad (18.52) \end{aligned}$$

By assumption, $n^{1/2} a\left(r_K^*\right) = O\left(n^{1/2}K^{-\alpha}\right) = o(1)$. By (18.14) and $nK^{-2\alpha} = o(1)$

$$\begin{aligned} n\gamma'_K\gamma_K &= n\mathbb{E}\left(r^*_{Ki}\boldsymbol{z}'_{Ki}\right)\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}'_{Ki}\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki} r^*_{Ki}\right) \\ &\leq nO\left(K^{-2\alpha}\right) \\ &= o(1). \end{aligned}$$

Together, both (18.51) and (18.52) are $o(1)$, as required. ■

Exercises

Exercise 18.1 You have a friend who wants to estimate β in the model

$$\begin{aligned} y_i &= x_i\beta + e_i \\ \mathbb{E}\left(e_i \mid z_i\right) &= 0 \end{aligned}$$

with both $x_i \in \mathbb{R}$ and $z_i \in \mathbb{R}$, and z_i is continuously distributed. Your friend wants to treat the reduced form equation for x_i as nonparametric

$$\begin{aligned} x_i &= g(z_i) + u_i \\ \mathbb{E}\left(u_i \mid z_i\right) &= 0 \end{aligned}$$

Your friend asks you for advice and help to construct an estimator $\widehat{\beta}$ of β. Describe an appropriate estimator. You do not have to develop the distribution theory, but try to be sufficiently complete with your advice so your friend can compute $\widehat{\beta}$.

Chapter 19

Empirical Likelihood

19.1 Non-Parametric Likelihood

An alternative to GMM is **empirical likelihood**. The idea is due to Art Owen (1988, 2001) and has been extended to moment condition models by Qin and Lawless (1994). It is a non-parametric analog of likelihood estimation.

The idea is to construct a multinomial distribution $F(p_1, ..., p_n)$ which places probability p_i at each observation. To be a valid multinomial distribution, these probabilities must satisfy the requirements that $p_i \geq 0$ and

$$\sum_{i=1}^{n} p_i = 1. \tag{19.1}$$

Since each observation is observed once in the sample, the log-likelihood function for this multinomial distribution is

$$\log L(p_1, ..., p_n) = \sum_{i=1}^{n} \log(p_i). \tag{19.2}$$

First let us consider a just-identified model. In this case the moment condition places no additional restrictions on the multinomial distribution. The maximum likelihood estimators of the probabilities $(p_1, ..., p_n)$ are those which maximize the log-likelihood subject to the constraint (19.1). This is equivalent to maximizing

$$\sum_{i=1}^{n} \log(p_i) - \mu \left(\sum_{i=1}^{n} p_i - 1 \right)$$

where μ is a Lagrange multiplier. The n first order conditions are $0 = p_i^{-1} - \mu$. Combined with the constraint (19.1) we find that the MLE is $p_i = n^{-1}$ yielding the log-likelihood $-n \log(n)$.

Now consider the case of an overidentified model with moment condition

$$\mathbb{E}(\boldsymbol{g}_i(\boldsymbol{\beta})) = \mathbf{0}$$

where $\boldsymbol{g}$ is $\ell \times 1$ and $\boldsymbol{\beta}$ is $k \times 1$ and for simplicity we write $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{g}(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i, \boldsymbol{\beta})$. The multinomial distribution which places probability p_i at each observation $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$ will satisfy this condition if and only if

$$\sum_{i=1}^{n} p_i \boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0} \tag{19.3}$$

The **empirical likelihood estimator** is the value of $\boldsymbol{\beta}$ which maximizes the multinomial log-likelihood (19.2) subject to the restrictions (19.1) and (19.3).

The Lagrangian for this maximization problem is

$$\mathcal{L}(\boldsymbol{\beta}, p_1, ..., p_n, \boldsymbol{\lambda}, \mu) = \sum_{i=1}^{n} \log(p_i) - \mu\left(\sum_{i=1}^{n} p_i - 1\right) - n\boldsymbol{\lambda}' \sum_{i=1}^{n} p_i \boldsymbol{g}_i(\boldsymbol{\beta})$$

where $\boldsymbol{\lambda}$ and μ are Lagrange multipliers. The first-order-conditions of $\mathcal{L}$ with respect to p_i, μ, and $\boldsymbol{\lambda}$ are

$$\begin{aligned} \frac{1}{p_i} &= \mu + n\boldsymbol{\lambda}' \boldsymbol{g}_i(\boldsymbol{\beta}) \\ \sum_{i=1}^{n} p_i &= 1 \\ \sum_{i=1}^{n} p_i \boldsymbol{g}_i(\boldsymbol{\beta}) &= \mathbf{0}. \end{aligned}$$

Multiplying the first equation by p_i, summing over i, and using the second and third equations, we find $\mu = n$ and

$$p_i = \frac{1}{n\left(1 + \boldsymbol{\lambda}' \boldsymbol{g}_i(\boldsymbol{\beta})\right)}.$$

Substituting into $\mathcal{L}$ we find

$$R(\boldsymbol{\beta}, \boldsymbol{\lambda}) = -n\log(n) - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}' \boldsymbol{g}_i(\boldsymbol{\beta})\right). \tag{19.4}$$

For given $\boldsymbol{\beta}$, the Lagrange multiplier $\boldsymbol{\lambda}(\boldsymbol{\beta})$ minimizes $R(\boldsymbol{\beta}, \boldsymbol{\lambda})$:

$$\boldsymbol{\lambda}(\boldsymbol{\beta}) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}}\, R(\boldsymbol{\beta}, \boldsymbol{\lambda}). \tag{19.5}$$

This minimization problem is the dual of the constrained maximization problem. The solution (when it exists) is well defined since $R(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is a convex function of $\boldsymbol{\lambda}$. The solution cannot be obtained explicitly, but must be obtained numerically (see section 6.5). This yields the (profile) empirical log-likelihood function for $\boldsymbol{\beta}$.

$$\begin{aligned} R(\boldsymbol{\beta}) &= R(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})) \\ &= -n\log(n) - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}(\boldsymbol{\beta})' \boldsymbol{g}_i(\boldsymbol{\beta})\right) \end{aligned}$$

The EL estimate $\widehat{\boldsymbol{\beta}}$ is the value which maximizes $R(\boldsymbol{\beta})$, or equivalently minimizes its negative

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left[-R(\boldsymbol{\beta})\right] \tag{19.6}$$

Numerical methods are required for calculation of $\widehat{\boldsymbol{\beta}}$ (see Section 19.5).

As a by-product of estimation, we also obtain the Lagrange multiplier $\widehat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\widehat{\boldsymbol{\beta}})$, probabilities

$$\widehat{p}_i = \frac{1}{n\left(1 + \widehat{\boldsymbol{\lambda}}' \boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right)\right)}.$$

and maximized empirical likelihood

$$R(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \log(\widehat{p}_i). \tag{19.7}$$

19.2 Asymptotic Distribution of EL Estimator

Define

$$\boldsymbol{G}_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\boldsymbol{\beta}) \tag{19.8}$$

$$\boldsymbol{G} = \mathbb{E}\left(\boldsymbol{G}_i(\boldsymbol{\beta})\right)$$

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\,\boldsymbol{g}_i(\boldsymbol{\beta})'\right)$$

and

$$\boldsymbol{V} = \left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1} \tag{19.9}$$

$$\boldsymbol{V}_{\boldsymbol{\lambda}} = \boldsymbol{\Omega} - \boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}' \tag{19.10}$$

For example, in the linear model, $\boldsymbol{G}_i(\boldsymbol{\beta}) = -\boldsymbol{z}_i\boldsymbol{x}_i'$, $\boldsymbol{G} = -\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{x}_i'\right)$, and $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)$.

Theorem 19.2.1 *Under regularity conditions,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right)$$

$$\sqrt{n}\widehat{\boldsymbol{\lambda}} \xrightarrow{d} \boldsymbol{\Omega}^{-1}\mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\lambda}}\right)$$

where $\boldsymbol{V}$ and $\boldsymbol{V}_{\boldsymbol{\lambda}}$ are defined in (19.9) and (19.10), and $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ and $\sqrt{n}\widehat{\boldsymbol{\lambda}}$ are asymptotically independent.

The theorem shows that the asymptotic variance $\boldsymbol{V}_{\boldsymbol{\beta}}$ for $\widehat{\boldsymbol{\beta}}$ is the same as for efficient GMM. Thus the EL estimator is asymptotically efficient.

Chamberlain (1987) showed that $\boldsymbol{V}_{\boldsymbol{\beta}}$ is the semiparametric efficiency bound for $\boldsymbol{\beta}$ in the overidentified moment condition model. This means that no consistent estimator for this class of models can have a lower asymptotic variance than $\boldsymbol{V}_{\boldsymbol{\beta}}$. Since the EL estimator achieves this bound, it is an asymptotically efficient estimator for $\boldsymbol{\beta}$.

Proof of Theorem 19.2.1. $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}})$ jointly solve

$$\boldsymbol{0} = \frac{\partial}{\partial \boldsymbol{\lambda}}R(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}}) = -\sum_{i=1}^{n}\frac{\boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right)}{\left(1 + \widehat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right)\right)} \tag{19.11}$$

$$\boldsymbol{0} = \frac{\partial}{\partial \boldsymbol{\beta}}R(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\lambda}}) = -\sum_{i=1}^{n}\frac{\boldsymbol{G}_i\left(\widehat{\boldsymbol{\beta}}\right)'\boldsymbol{\lambda}}{1 + \widehat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right)}. \tag{19.12}$$

Let $\boldsymbol{G}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{G}_i(\boldsymbol{\beta})$, $\overline{\boldsymbol{g}}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta})$ and $\boldsymbol{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta})\,\boldsymbol{g}_i(\boldsymbol{\beta})'$.

Expanding (19.12) around $\boldsymbol{\beta}$ and $\boldsymbol{\lambda} = \boldsymbol{0}$ yields

$$\boldsymbol{0} \simeq \boldsymbol{G}_n'\widehat{\boldsymbol{\lambda}}. \tag{19.13}$$

Expanding (19.11) around $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 = \boldsymbol{0}$ yields

$$\boldsymbol{0} \simeq -\overline{\boldsymbol{g}}_n - \boldsymbol{G}_n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \boldsymbol{\Omega}_n\widehat{\boldsymbol{\lambda}} \tag{19.14}$$

Premultiplying by $\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}$ and using (19.13) yields

$$\begin{aligned}\mathbf{0} &\simeq -\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\overline{\boldsymbol{g}}_n - \boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) + \boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{\Omega}_n\widehat{\boldsymbol{\lambda}} \\ &= -\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\overline{\boldsymbol{g}}_n - \boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\end{aligned}$$

Solving for $\widehat{\boldsymbol{\beta}}$ and using the WLLN and CLT yields

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &\simeq -\left(\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\sqrt{n}\overline{\boldsymbol{g}}_n \\ &\xrightarrow{d} \left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right) \\ &= \mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right)\end{aligned} \tag{19.15}$$

Solving (19.14) for $\widehat{\boldsymbol{\lambda}}$ and using (19.15) yields

$$\begin{aligned}\sqrt{n}\widehat{\boldsymbol{\lambda}} &\simeq \boldsymbol{\Omega}_n^{-1}\left(\boldsymbol{I}-\boldsymbol{G}_n\left(\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\right)\sqrt{n}\overline{\boldsymbol{g}}_n \\ &\xrightarrow{d} \boldsymbol{\Omega}^{-1}\left(\boldsymbol{I}-\boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\right)\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right) \\ &= \boldsymbol{\Omega}^{-1}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\lambda}}\right)\end{aligned} \tag{19.16}$$

Furthermore, since

$$\boldsymbol{G}'\left(\boldsymbol{I}-\boldsymbol{\Omega}^{-1}\boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\right)=\mathbf{0}$$

$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)$ and $\sqrt{n}\widehat{\boldsymbol{\lambda}}$ are asymptotically uncorrelated and hence independent.

19.3 Overidentifying Restrictions

In a parametric likelihood context, tests are based on the difference in the log likelihood functions. The same statistic can be constructed for empirical likelihood. Twice the difference between the unrestricted empirical log-likelihood $-n\log(n)$ and the maximized empirical log-likelihood for the model (19.7) is

$$LR_n = \sum_{i=1}^{n} 2\log\left(1+\widehat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right)\right). \tag{19.17}$$

Theorem 19.3.1 *If* $\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\right)=\mathbf{0}$ *then* $LR_n \xrightarrow{d} \chi^2_{\ell-k}$.

The EL overidentification test is similar to the GMM overidentification test. They are asymptotically first-order equivalent, and have the same interpretation. The overidentification test is a very useful by-product of EL estimation, and it is advisable to report the statistic LR_n whenever EL is the estimation method.

Proof of Theorem 19.3.1. First, by a Taylor expansion, (19.15), and (19.16),

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right) &\simeq \sqrt{n}\left(\overline{\boldsymbol{g}}_n+\boldsymbol{G}_n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right) \\ &\simeq \left(\boldsymbol{I}-\boldsymbol{G}_n\left(\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\right)\sqrt{n}\overline{\boldsymbol{g}}_n \\ &\simeq \boldsymbol{\Omega}_n\sqrt{n}\widehat{\boldsymbol{\lambda}}.\end{aligned}$$

Second, since $\log(1+u) \simeq u - u^2/2$ for u small,

$$\begin{aligned} LR_n &= \sum_{i=1}^{n} 2\log\left(1 + \widehat{\boldsymbol{\lambda}}' \boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right)\right) \\ &\simeq 2\widehat{\boldsymbol{\lambda}}' \sum_{i=1}^{n} \boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right) - \widehat{\boldsymbol{\lambda}}' \sum_{i=1}^{n} \boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right) \boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right)' \widehat{\boldsymbol{\lambda}} \\ &\simeq n\widehat{\boldsymbol{\lambda}}' \boldsymbol{\Omega}_n \widehat{\boldsymbol{\lambda}} \\ &\xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\lambda}}\right)' \boldsymbol{\Omega}^{-1} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\lambda}}\right) \\ &= \chi^2_{\ell-k} \end{aligned}$$

where the proof of the final equality is left as an exercise.

19.4 Testing

Let the maintained model be

$$\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\right) = \mathbf{0} \tag{19.18}$$

where $\boldsymbol{g}$ is $\ell \times 1$ and $\boldsymbol{\beta}$ is $k \times 1$. By "maintained" we mean that the overidentfying restrictions contained in (19.18) are assumed to hold and are not being challenged (at least for the test discussed in this section). The hypothesis of interest is

$$\boldsymbol{h}(\boldsymbol{\beta}) = \mathbf{0}.$$

where $\boldsymbol{h} : \mathbb{R}^k \to \mathbb{R}^a$. The restricted EL estimator and likelihood are the values which solve

$$\begin{aligned} \widetilde{\boldsymbol{\beta}} &= \underset{\boldsymbol{h}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmax}} R(\boldsymbol{\beta}) \\ R(\widetilde{\boldsymbol{\beta}}) &= \max_{\boldsymbol{h}(\boldsymbol{\beta})=\mathbf{0}} R(\boldsymbol{\beta}). \end{aligned}$$

Fundamentally, the restricted EL estimator $\tilde{\boldsymbol{\beta}}$ is simply an EL estimator with $\ell - k + a$ overidentifying restrictions, so there is no fundamental change in the distribution theory for $\widetilde{\boldsymbol{\beta}}$ relative to $\widehat{\boldsymbol{\beta}}$. To test the hypothesis $\boldsymbol{h}(\boldsymbol{\beta})$ while maintaining (19.18), the simple overidentifying restrictions test (19.17) is not appropriate. Instead we use the difference in log-likelihoods:

$$LR_n = 2\left(R(\widehat{\boldsymbol{\beta}}) - R(\widetilde{\boldsymbol{\beta}})\right).$$

This test statistic is a natural analog of the GMM distance statistic.

Theorem 19.4.1 *Under (19.18) and* $\mathbb{H}_0 : \boldsymbol{h}(\boldsymbol{\beta}) = \mathbf{0}$, $LR_n \xrightarrow{d} \chi^2_a$.

The proof of this result is more challenging and is omitted.

19.5 Numerical Computation

Derivatives

The numerical calculations depend on derivatives of the dual likelihood function (19.4). Define

$$\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{\boldsymbol{g}_i(\boldsymbol{\beta})}{\left(1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})\right)}$$
$$\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{\boldsymbol{G}_i(\boldsymbol{\beta})'\boldsymbol{\lambda}}{1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})}$$

The first derivatives of (19.4) are

$$\boldsymbol{R}_{\boldsymbol{\lambda}} = \frac{\partial}{\partial\boldsymbol{\lambda}}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = -\sum_{i=1}^{n}\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})$$
$$\boldsymbol{R}_{\boldsymbol{\beta}} = \frac{\partial}{\partial\boldsymbol{\beta}}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = -\sum_{i=1}^{n}\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda}).$$

The second derivatives are

$$\boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} = \frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^{n}\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})\,\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})'$$
$$\boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}} = \frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\beta}'}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^{n}\left(\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})\,\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})' - \frac{\boldsymbol{G}_i(\boldsymbol{\beta})}{1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})}\right)$$
$$\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^{n}\left(\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})\,\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})' - \frac{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\left(\boldsymbol{g}_i(\boldsymbol{\beta})'\boldsymbol{\lambda}\right)}{1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})}\right)$$

Inner Loop

The so-called "inner loop" solves (19.5) for given $\boldsymbol{\beta}$. The modified Newton method takes a quadratic approximation to $R_n(\boldsymbol{\beta},\boldsymbol{\lambda})$ yielding the iteration rule

$$\boldsymbol{\lambda}_{j+1} = \boldsymbol{\lambda}_j - \delta\left(\boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j)\right)^{-1}\boldsymbol{R}_{\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j). \tag{19.19}$$

where $\delta > 0$ is a scalar steplength (to be discussed next). The starting value $\boldsymbol{\lambda}_1$ can be set to the zero vector. The iteration (19.19) is continued until the gradient $R_{\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j)$ is smaller than some prespecified tolerance.

Efficient convergence requires a good choice of steplength δ. One method uses the following quadratic approximation. Set $\delta_0 = 0$, $\delta_1 = \frac{1}{2}$ and $\delta_2 = 1$. For $p = 0, 1, 2$, set

$$\boldsymbol{\lambda}_p = \boldsymbol{\lambda}_j - \delta_p\left(\boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j)\right)^{-1}\boldsymbol{R}_{\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j))$$
$$R_p = R(\boldsymbol{\beta},\boldsymbol{\lambda}_p)$$

A quadratic function can be fit exactly through these three points. The value of δ which minimizes this quadratic is

$$\hat{\delta} = \frac{R_2 + 3R_0 - 4R_1}{4R_2 + 4R_0 - 8R_1}.$$

yielding the steplength to be plugged into (19.19).

A complication is that $\boldsymbol{\lambda}$ must be constrained so that $0 \le p_i \le 1$ which holds if

$$n\left(1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})\right) \ge 1 \tag{19.20}$$

for all i. If (19.20) fails, the stepsize δ needs to be decreased.

Outer Loop

The outer loop is the minimization (19.6). This can be done by the modified Newton method described in the previous section. The gradient for (19.6) is

$$\boldsymbol{R}_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} R(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} R(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \boldsymbol{R}_{\boldsymbol{\beta}} + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}} = \boldsymbol{R}_{\boldsymbol{\beta}}$$

since $\boldsymbol{R}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = 0$ at $\boldsymbol{\lambda} = \boldsymbol{\lambda}(\boldsymbol{\beta})$, where

$$\boldsymbol{\lambda}_{\boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}'} \boldsymbol{\lambda}(\boldsymbol{\beta}) = -\boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1} \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}},$$

the second equality following from the implicit function theorem applied to $\boldsymbol{R}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})) = 0$.

The Hessian for (19.6) is

$$\begin{aligned}
\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} R(\boldsymbol{\beta}) \\
&= -\frac{\partial}{\partial \boldsymbol{\beta}'} \left[\boldsymbol{R}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})) + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}}(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})) \right] \\
&= -\left(\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})) + \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}}' \boldsymbol{\lambda}_{\boldsymbol{\beta}} + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}} + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \boldsymbol{\lambda}_{\boldsymbol{\beta}} \right) \\
&= \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1} \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}} - \boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}.
\end{aligned}$$

It is not guaranteed that $\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}} > 0$. If not, the eigenvalues of $\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}$ should be adjusted so that all are positive. The Newton iteration rule is

$$\boldsymbol{\beta}_{j+1} = \boldsymbol{\beta}_j - \delta \boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}^{-1} \boldsymbol{R}_{\boldsymbol{\beta}}$$

where δ is a scalar stepsize, and the rule is iterated until convergence.

Chapter 20

Regression Extensions

20.1 Nonlinear Least Squares

In some cases we might use a parametric regression function $m(\boldsymbol{x}, \boldsymbol{\theta}) = \mathbb{E}(y_i \mid \boldsymbol{x}_i = \boldsymbol{x})$ which is a non-linear function of the parameters $\boldsymbol{\theta}$. We describe this setting as **nonlinear regression**.

Example 20.1.1 *Exponential Link Regression*

$$m(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left(\boldsymbol{x}'\boldsymbol{\theta}\right)$$

The exponential link function is strictly positive, so this choice can be useful when it is desired to constrain the mean to be strictly positive.

Example 20.1.2 *Logistic Link Regression*

$$m(\boldsymbol{x}, \boldsymbol{\theta}) = \Lambda\left(\boldsymbol{x}'\boldsymbol{\theta}\right)$$

where

$$\Lambda(u) = (1 + \exp(-u))^{-1} \tag{20.1}$$

is the Logistic distribution function. Since the logistic link function lies in $[0, 1]$*, this choice can be useful when the conditional mean is bounded between 0 and 1.*

Example 20.1.3 *Exponentially Transformed Regressors*

$$m(x, \boldsymbol{\theta}) = \theta_1 + \theta_2 \exp(\theta_3 x)$$

Example 20.1.4 *Power Transformation*

$$m(x, \boldsymbol{\theta}) = \theta_1 + \theta_2 x^{\theta_3}$$

with $x > 0$.

Example 20.1.5 *Box-Cox Transformed Regressors*

$$m(x, \boldsymbol{\theta}) = \theta_1 + \theta_2 x^{(\theta_3)}$$

where

$$x^{(\lambda)} = \left\{ \begin{array}{ll} \dfrac{x^\lambda - 1}{\lambda}, & \text{if } \lambda > 0 \\ \log(x), & \text{if } \lambda = 0 \end{array} \right\} \tag{20.2}$$

and $x > 0$*. The function (20.2) is called the Box-Cox Transformation and was introduced by Box and Cox (1964). The function nests linearity* ($\lambda = 1$) *and logarithmic* ($\lambda = 0$) *transformations continuously.*

Example 20.1.6 *Continuous Threshold Regression*

$$m(x, \boldsymbol{\theta}) = \theta_1 + \theta_2 x + \theta_3 (x - \theta_4) 1(x > \theta_4)$$

Example 20.1.7 *Threshold Regression*

$$m(\boldsymbol{x}, \boldsymbol{\theta}) = \left(\theta_1' \boldsymbol{x}_1\right) 1(x_2 < \theta_3) + \left(\theta_2' \boldsymbol{x}_1\right) 1(x_2 \geq \theta_3)$$

Example 20.1.8 *Smooth Transition*

$$m(\boldsymbol{x}, \boldsymbol{\theta}) = \theta_1' \boldsymbol{x}_1 + \left(\theta_2' \boldsymbol{x}_1\right) \Lambda\left(\frac{x_2 - \theta_3}{\theta_4}\right)$$

where $\Lambda(u)$ *is the logit function (20.1).*

What differentiates these examples from the linear regression model is that the conditional mean cannot be written as a linear function of the parameter vector $\boldsymbol{\theta}$.

Nonlinear regression is sometimes adopted because the functional form $m(\boldsymbol{x}, \boldsymbol{\theta})$ is suggested by an economic model. In other cases, it is adopted as a flexible approximation to an unknown regression function.

The least squares estimator $\widehat{\boldsymbol{\theta}}$ minimizes the normalized sum-of-squared-errors

$$\widehat{S}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta}))^2 .$$

When the regression function is nonlinear, we call $\widehat{\boldsymbol{\theta}}$ the **nonlinear least squares** (NLLS) estimator. The NLLS residuals are $\widehat{e}_i = y_i - m\left(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}\right)$.

One motivation for the choice of NLLS as the estimation method is that the parameter $\boldsymbol{\theta}$ is the solution to the population problem $\min_{\boldsymbol{\theta}} \mathbb{E}(y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta}))^2$

Since the criterion $\widehat{S}(\boldsymbol{\theta})$ is not quadratic, $\widehat{\boldsymbol{\theta}}$ must be found by numerical methods. See Appendix E. When $m(\boldsymbol{x}, \boldsymbol{\theta})$ is differentiable, then the FOC for minimization are

$$\mathbf{0} = \sum_{i=1}^{n} \boldsymbol{m}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}}\right) \widehat{e}_i \tag{20.3}$$

where

$$\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} m(\boldsymbol{x}, \boldsymbol{\theta}) .$$

Theorem 20.1.1 ***Asymptotic Distribution of NLLS Estimator***
If the model is identified and $m(\boldsymbol{x}, \boldsymbol{\theta})$ *is differentiable with respect to* $\boldsymbol{\theta}$*,*

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\theta}})$$

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \left(\mathbb{E}\left(\boldsymbol{m}_{\boldsymbol{\theta}i} \boldsymbol{m}_{\boldsymbol{\theta}i}'\right)\right)^{-1} \left(\mathbb{E}\left(\boldsymbol{m}_{\boldsymbol{\theta}i} \boldsymbol{m}_{\boldsymbol{\theta}i}' e_i^2\right)\right) \left(\mathbb{E}\left(\boldsymbol{m}_{\boldsymbol{\theta}i} \boldsymbol{m}_{\boldsymbol{\theta}i}'\right)\right)^{-1}$$

where $\boldsymbol{m}_{\boldsymbol{\theta}i} = \boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{\theta}_0)$.

Based on Theorem 20.1.1, an estimate of the asymptotic variance $\boldsymbol{V}_{\boldsymbol{\theta}}$ is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{m}}_{\boldsymbol{\theta}i}\widehat{\boldsymbol{m}}_{\boldsymbol{\theta}i}'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{m}}_{\boldsymbol{\theta}i}\widehat{\boldsymbol{m}}_{\boldsymbol{\theta}i}'\widehat{e}_i^2\right) \left(\frac{1}{n}\sum_{i=1}^{n} \widehat{\boldsymbol{m}}_{\boldsymbol{\theta}i}\widehat{\boldsymbol{m}}_{\boldsymbol{\theta}i}'\right)^{-1}$$

where $\widehat{\boldsymbol{m}}_{\boldsymbol{\theta}i} = \boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})$ and $\widehat{e}_i = y_i - m(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})$.

Identification is often tricky in nonlinear regression models. Suppose that

$$m(\boldsymbol{x}_i, \boldsymbol{\theta}) = \boldsymbol{\beta}_1'\boldsymbol{z}_i + \boldsymbol{\beta}_2'\boldsymbol{x}_i(\gamma)$$

where $\boldsymbol{x}_i(\gamma)$ is a function of $\boldsymbol{x}_i$ and the unknown parameter $\boldsymbol{\gamma}$. Examples include $x_i(\gamma) = x_i^{\gamma}$, $x_i(\gamma) = \exp(\gamma x_i)$, and $x_i(\boldsymbol{\gamma}) = x_i 1(g(x_i) > \gamma)$. The model is linear when $\boldsymbol{\beta}_2 = \boldsymbol{0}$, and this is often a useful hypothesis (sub-model) to consider. Thus we want to test

$$\mathbb{H}_0 : \boldsymbol{\beta}_2 = \boldsymbol{0}.$$

However, under $\mathbb{H}_0$, the model is

$$y_i = \boldsymbol{\beta}_1'\boldsymbol{z}_i + e_i$$

and both $\boldsymbol{\beta}_2$ and γ have dropped out. This means that under $\mathbb{H}_0$, γ is not identified. This renders the distribution theory presented in the previous section invalid. Thus when the truth is that $\boldsymbol{\beta}_2 = \boldsymbol{0}$, the parameter estimates are not asymptotically normally distributed. Furthermore, tests of $\mathbb{H}_0$ do not have asymptotic normal or chi-square distributions.

The asymptotic theory of such tests have been worked out by Andrews and Ploberger (1994) and B. E. Hansen (1996). In particular, Hansen shows how to use simulation (similar to the bootstrap) to construct the asymptotic critical values (or p-values) in a given application.

Proof of Theorem 20.1.1 (Sketch). NLLS estimation falls in the class of optimization estimators. For this theory, it is useful to denote the true value of the parameter $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_0$.

The first step is to show that $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Proving that nonlinear estimators are consistent is more challenging than for linear estimators. We sketch the main argument. The idea is that $\widehat{\boldsymbol{\theta}}$ minimizes the sample criterion function $\widehat{S}(\boldsymbol{\theta})$, which (for any $\boldsymbol{\theta}$) converges in probability to the mean-squared error function $\mathbb{E}\left((y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta}))^2\right)$. Thus it seems reasonable that the minimizer $\widehat{\boldsymbol{\theta}}$ will converge in probability to $\boldsymbol{\theta}_0$, the minimizer of $\mathbb{E}\left((y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta}))^2\right)$. It turns out that to show this rigorously, we need to show that $\widehat{S}(\boldsymbol{\theta})$ converges *uniformly* to its expectation $\mathbb{E}\left((y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta}))^2\right)$, which means that the maximum discrepancy must converge in probability to zero, to exclude the possibility that $\widehat{S}(\boldsymbol{\theta})$ is excessively wiggly in $\boldsymbol{\theta}$. Proving uniform convergence is technically challenging, but it can be shown to hold broadly for relevant nonlinear regression models, especially if the regression function $m(\boldsymbol{x}_i, \boldsymbol{\theta})$ is differentiable in $\boldsymbol{\theta}$. For a complete treatment of the theory of optimization estimators see Newey and McFadden (1994).

Since $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, $\widehat{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}_0$ for n large, so the minimization of $\widehat{S}(\boldsymbol{\theta})$ only needs to be examined for $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_0$. Let

$$y_i^0 = e_i + \boldsymbol{m}_{\boldsymbol{\theta}i}'\boldsymbol{\theta}_0.$$

For $\boldsymbol{\theta}$ close to the true value $\boldsymbol{\theta}_0$, by a first-order Taylor series approximation,

$$m(\boldsymbol{x}_i, \boldsymbol{\theta}) \simeq m(\boldsymbol{x}_i, \boldsymbol{\theta}_0) + \boldsymbol{m}_{\boldsymbol{\theta}i}'(\boldsymbol{\theta} - \boldsymbol{\theta}_0).$$

Thus

$$\begin{aligned} y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta}) &\simeq (e_i + m(\boldsymbol{x}_i, \boldsymbol{\theta}_0)) - \left(m(\boldsymbol{x}_i, \boldsymbol{\theta}_0) + \boldsymbol{m}_{\boldsymbol{\theta}i}'(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right) \\ &= e_i - \boldsymbol{m}_{\boldsymbol{\theta}i}'(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= y_i^0 - \boldsymbol{m}_{\boldsymbol{\theta}i}'\boldsymbol{\theta}. \end{aligned}$$

Hence the normalized sum of squared errors function is

$$\widehat{S}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)^2 \simeq \frac{1}{n}\sum_{i=1}^{n}\left(y_i^0 - \boldsymbol{m}_{\boldsymbol{\theta}i}'\boldsymbol{\theta}\right)^2$$

and the right-hand-side is the criterion function for a linear regression of y_i^0 on $\boldsymbol{m}_{\boldsymbol{\theta}i}$. Thus the NLLS estimator $\widehat{\boldsymbol{\theta}}$ has the same asymptotic distribution as the (infeasible) OLS regression of y_i^0 on $\boldsymbol{m}_{\boldsymbol{\theta}i}$, which is that stated in the theorem.

20.2 Generalized Least Squares

In the projection model, we know that the least-squares estimator is semi-parametrically efficient for the projection coefficient. However, in the linear regression model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0, \end{aligned}$$

the least-squares estimator is inefficient. The theory of Chamberlain (1987) can be used to show that in this model the semiparametric efficiency bound is obtained by the **Generalized Least Squares** (GLS) estimator (4.19) introduced in Section 4.7.1. The GLS estimator is sometimes called the Aitken estimator. The GLS estimator (20.2) is infeasible since the matrix $\boldsymbol{D}$ is unknown. A feasible GLS (FGLS) estimator replaces the unknown $\boldsymbol{D}$ with an estimate $\widehat{\boldsymbol{D}} = \text{diag}\{\widehat{\sigma}_1^2, ..., \widehat{\sigma}_n^2\}$. We now discuss this estimation problem.

Suppose that we model the conditional variance using the parametric form

$$\begin{aligned} \sigma_i^2 &= \alpha_0 + \boldsymbol{z}_{1i}'\boldsymbol{\alpha}_1 \\ &= \boldsymbol{\alpha}'\boldsymbol{z}_i, \end{aligned}$$

where $\boldsymbol{z}_{1i}$ is some $q \times 1$ function of $\boldsymbol{x}_i$. Typically, $\boldsymbol{z}_{1i}$ are squares (and perhaps levels) of some (or all) elements of $\boldsymbol{x}_i$. Often the functional form is kept simple for parsimony.

Let $\eta_i = e_i^2$. Then

$$\mathbb{E}\left(\eta_i \mid \boldsymbol{x}_i\right) = \alpha_0 + \boldsymbol{z}_{1i}'\boldsymbol{\alpha}_1$$

and we have the regression equation

$$\begin{aligned} \eta_i &= \alpha_0 + \boldsymbol{z}_{1i}'\boldsymbol{\alpha}_1 + \xi_i \\ \mathbb{E}\left(\xi_i \mid \boldsymbol{x}_i\right) &= 0. \end{aligned} \tag{20.4}$$

This regression error ξ_i is generally heteroskedastic and has the conditional variance

$$\begin{aligned} \text{var}\left(\xi_i \mid \boldsymbol{x}_i\right) &= \text{var}\left(e_i^2 \mid \boldsymbol{x}_i\right) \\ &= \mathbb{E}\left(\left(e_i^2 - \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right)\right)^2 \mid \boldsymbol{x}_i\right) \\ &= \mathbb{E}\left(e_i^4 \mid \boldsymbol{x}_i\right) - \left(\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right)\right)^2. \end{aligned}$$

Suppose e_i (and thus η_i) were observed. Then we could estimate $\boldsymbol{\alpha}$ by OLS:

$$\widehat{\boldsymbol{\alpha}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{\eta} \xrightarrow{p} \boldsymbol{\alpha}$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) \xrightarrow{d} \text{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\alpha}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\alpha}} = \left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\xi_i^2\right)\left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}. \tag{20.5}$$

While e_i is not observed, we have the OLS residual $\widehat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} = e_i - \boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Thus

$$\begin{aligned}\phi_i &\equiv \widehat{\eta}_i - \eta_i \\ &= \widehat{e}_i^2 - e_i^2 \\ &= -2e_i\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{x}_i\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).\end{aligned}$$

And then

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{z}_i\phi_i &= \frac{-2}{n}\sum_{i=1}^n \boldsymbol{z}_i e_i\boldsymbol{x}_i'\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{x}_i\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\sqrt{n} \\ &\xrightarrow{p} \mathbf{0}\end{aligned}$$

Let

$$\widetilde{\boldsymbol{\alpha}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\widehat{\boldsymbol{\eta}} \tag{20.6}$$

be from OLS regression of $\widehat{\eta}_i$ on $\boldsymbol{z}_i$. Then

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) &= \sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) + \left(n^{-1}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} n^{-1/2}\boldsymbol{Z}'\boldsymbol{\phi} \\ &\xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\alpha}}\right)\end{aligned} \tag{20.7}$$

Thus the fact that η_i is replaced with $\widehat{\eta}_i$ is asymptotically irrelevant. We call (20.6) the *skedastic* regression, as it is estimating the conditional variance of the regression of y_i on $\boldsymbol{x}_i$. We have shown that $\boldsymbol{\alpha}$ is consistently estimated by a simple procedure, and hence we can estimate $\sigma_i^2 = \boldsymbol{z}_i'\boldsymbol{\alpha}$ by

$$\widetilde{\sigma}_i^2 = \widetilde{\boldsymbol{\alpha}}'\boldsymbol{z}_i. \tag{20.8}$$

Suppose that $\widetilde{\sigma}_i^2 > 0$ for all i. Then set

$$\widetilde{\boldsymbol{D}} = \mathrm{diag}\{\widetilde{\sigma}_1^2, ..., \widetilde{\sigma}_n^2\}$$

and

$$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{y}.$$

This is the feasible GLS, or FGLS, estimator of $\boldsymbol{\beta}$. Since there is not a unique specification for the conditional variance the FGLS estimator is not unique, and will depend on the model (and estimation method) for the skedastic regression.

One typical problem with implementation of FGLS estimation is that in the linear specification (20.4), there is no guarantee that $\widetilde{\sigma}_i^2 > 0$ for all i. If $\widetilde{\sigma}_i^2 < 0$ for some i, then the FGLS estimator is not well defined. Furthermore, if $\widetilde{\sigma}_i^2 \approx 0$ for some i then the FGLS estimator will force the regression equation through the point $(y_i, \boldsymbol{x}_i)$, which is undesirable. This suggests that there is a need to bound the estimated variances away from zero. A trimming rule takes the form

$$\overline{\sigma}_i^2 = \max[\widetilde{\sigma}_i^2, c\widehat{\sigma}^2]$$

for some $c > 0$. For example, setting $c = 1/4$ means that the conditional variance function is constrained to exceed one-fourth of the unconditional variance. As there is no clear method to select c, this introduces a degree of arbitrariness. In this context it is useful to re-estimate the model with several choices for the trimming parameter. If the estimates turn out to be sensitive to its choice, the estimation method should probably be reconsidered.

It is possible to show that if the skedastic regression is correctly specified, then FGLS is asymptotically equivalent to GLS. As the proof is tricky, we just state the result without proof.

Theorem 20.2.1 *If the skedastic regression is correctly specified,*

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{GLS}-\widetilde{\boldsymbol{\beta}}_{FGLS}\right)\xrightarrow{p}\mathbf{0},$$

and thus

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{FGLS}-\boldsymbol{\beta}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right),$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}}=\left(\mathbb{E}\left(\sigma_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}.$$

Examining the asymptotic distribution of Theorem 20.2.1, the natural estimator of the asymptotic variance of $\widetilde{\boldsymbol{\beta}}$ is

$$\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}=\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\sigma}_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}=\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}.$$

which is consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}$ as $n\to\infty$. This estimator $\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}$ is appropriate when the skedastic regression (20.4) is correctly specified.

It may be the case that $\boldsymbol{\alpha}'\boldsymbol{z}_i$ is only an approximation to the true conditional variance $\sigma_i^2=\mathbb{E}(e_i^2\mid\boldsymbol{x}_i)$. In this case we interpret $\boldsymbol{\alpha}'\boldsymbol{z}_i$ as a linear projection of e_i^2 on $\boldsymbol{z}_i$. $\widetilde{\boldsymbol{\beta}}$ should perhaps be called a quasi-FGLS estimator of $\boldsymbol{\beta}$. Its asymptotic variance is not that given in Theorem 20.2.1. Instead,

$$\boldsymbol{V}_{\boldsymbol{\beta}}=\left(\mathbb{E}\left(\left(\boldsymbol{\alpha}'\boldsymbol{z}_i\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}\left(\mathbb{E}\left(\left(\boldsymbol{\alpha}'\boldsymbol{z}_i\right)^{-2}\sigma_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)\left(\mathbb{E}\left(\left(\boldsymbol{\alpha}'\boldsymbol{z}_i\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}.$$

$\boldsymbol{V}_{\boldsymbol{\beta}}$ takes a sandwich form similar to the covariance matrix of the OLS estimator. Unless $\sigma_i^2=\boldsymbol{\alpha}'\boldsymbol{z}_i$, $\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}$ is inconsistent for $\boldsymbol{V}_{\boldsymbol{\beta}}$.

An appropriate solution is to use a White-type estimator in place of $\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0}$. This may be written as

$$\begin{aligned}\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}&=\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\sigma}_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\sigma}_i^{-4}\widehat{e}_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\frac{1}{n}\sum_{i=1}^{n}\widetilde{\sigma}_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\\&=\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{D}}\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}\end{aligned}$$

where $\widehat{\boldsymbol{D}}=\mathrm{diag}\{\widehat{e}_1^2,...,\widehat{e}_n^2\}$. This is estimator is robust to misspecification of the conditional variance, and was proposed by Cragg (1992).

In the linear regression model, FGLS is asymptotically superior to OLS. Why then do we not exclusively estimate regression models by FGLS? This is a good question. There are three reasons.

First, FGLS estimation depends on specification and estimation of the skedastic regression. Since the form of the skedastic regression is unknown, and it may be estimated with considerable error, the estimated conditional variances may contain more noise than information about the true conditional variances. In this case, FGLS can do worse than OLS in practice.

Second, individual estimated conditional variances may be negative, and this requires trimming to solve. This introduces an element of arbitrariness which is unsettling to empirical researchers.

Third, and probably most importantly, OLS is a robust estimator of the parameter vector. It is consistent not only in the regression model, but also under the assumptions of linear projection. The GLS and FGLS estimators, on the other hand, require the assumption of a correct conditional

mean. If the equation of interest is a linear projection and not a conditional mean, then the OLS and FGLS estimators will converge in probability to different limits as they will be estimating two different projections. The FGLS probability limit will depend on the particular function selected for the skedastic regression. The point is that the efficiency gains from FGLS are built on the stronger assumption of a correct conditional mean, and the cost is a loss of robustness to misspecification.

20.3 Testing for Heteroskedasticity

The hypothesis of homoskedasticity is that $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, or equivalently that

$$\mathbb{H}_0 : \boldsymbol{\alpha}_1 = 0$$

in the regression (20.4). We may therefore test this hypothesis by the estimation (20.6) and constructing a Wald statistic. In the classic literature it is typical to impose the stronger assumption that e_i is independent of $\boldsymbol{x}_i$, in which case ξ_i is independent of $\boldsymbol{x}_i$ and the asymptotic variance (20.5) for $\tilde{\boldsymbol{\alpha}}$ simplifies to

$$V_{\boldsymbol{\alpha}} = \left(\mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i'\right)\right)^{-1} \mathbb{E}\left(\xi_i^2\right). \tag{20.9}$$

Hence the standard test of $\mathbb{H}_0$ is a classic F (or Wald) test for exclusion of all regressors from the skedastic regression (20.6). The asymptotic distribution (20.7) and the asymptotic variance (20.9) under independence show that this test has an asymptotic chi-square distribution.

Theorem 20.3.1 *Under $\mathbb{H}_0$ and e_i independent of $\boldsymbol{x}_i$, the Wald test of $\mathbb{H}_0$ is asymptotically χ_q^2.*

Most tests for heteroskedasticity take this basic form. The main differences between popular tests are which transformations of $\boldsymbol{x}_i$ enter $\boldsymbol{z}_i$. Motivated by the form of the asymptotic variance of the OLS estimator $\widehat{\boldsymbol{\beta}}$, White (1980) proposed that the test for heteroskedasticity be based on setting $\boldsymbol{z}_i$ to equal all non-redundant elements of $\boldsymbol{x}_i$, its squares, and all cross-products. Breusch-Pagan (1979) proposed what might appear to be a distinct test, but the only difference is that they allowed for general choice of $\boldsymbol{z}_i$, and replaced $\mathbb{E}\left(\xi_i^2\right)$ with $2\sigma^4$ which holds when e_i is $\mathrm{N}\left(\mathbf{0}, \sigma^2\right)$. If this simplification is replaced by the standard formula (under independence of the error), the two tests coincide.

It is important not to misuse tests for heteroskedasticity. It should not be used to determine whether to estimate a regression equation by OLS or FGLS, nor to determine whether classic or White standard errors should be reported. Hypothesis tests are not designed for these purposes. Rather, tests for heteroskedasticity should be used to answer the scientific question of whether or not the conditional variance is a function of the regressors. If this question is not of economic interest, then there is no value in conducting a test for heteorskedasticity

20.4 Testing for Omitted Nonlinearity

If the goal is to estimate the conditional expectation $\mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right)$, it is useful to have a general test of the adequacy of the specification.

One simple test for neglected nonlinearity is to add nonlinear functions of the regressors to the regression, and test their significance using a Wald test. Thus, if the model $y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \widehat{e}_i$ has been fit by OLS, let $\boldsymbol{z}_i = \boldsymbol{h}(\boldsymbol{x}_i)$ denote functions of $\boldsymbol{x}_i$ which are not linear functions of $\boldsymbol{x}_i$ (perhaps squares of non-binary regressors) and then fit $y_i = \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}} + \boldsymbol{z}_i'\widetilde{\boldsymbol{\gamma}} + \widetilde{e}_i$ by OLS, and form a Wald statistic for $\boldsymbol{\gamma} = \mathbf{0}$.

Another popular approach is the RESET test proposed by Ramsey (1969). The null model is

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$$

which is estimated by OLS, yielding predicted values $\widehat{y}_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$. Now let

$$\boldsymbol{z}_i = \begin{pmatrix} \widehat{y}_i^2 \\ \vdots \\ \widehat{y}_i^m \end{pmatrix}$$

be a $(m-1)$-vector of powers of $\widehat{y}_i$. Then run the auxiliary regression

$$y_i = \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}} + \boldsymbol{z}_i'\widetilde{\boldsymbol{\gamma}} + \widetilde{e}_i \tag{20.10}$$

by OLS, and form the Wald statistic W for $\boldsymbol{\gamma} = \boldsymbol{0}$. It is easy (although somewhat tedious) to show that under the null hypothesis, $W \xrightarrow{d} \chi^2_{m-1}$. Thus the null is rejected at the α% level if W exceeds the upper $1-\alpha$ critical value of the χ^2_{m-1} distribution.

To implement the test, m must be selected in advance. Typically, small values such as $m = 2$, 3, or 4 seem to work best.

The RESET test appears to work well as a test of functional form against a wide range of smooth alternatives. It is particularly powerful at detecting *single-index* models of the form

$$y_i = G(\boldsymbol{x}_i'\boldsymbol{\beta}) + e_i$$

where $G(\cdot)$ is a smooth "link" function. To see why this is the case, note that (20.10) may be written as

$$y_i = \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}} + \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right)^2\widetilde{\gamma}_1 + \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right)^3\widetilde{\gamma}_2 + \cdots \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right)^m\widetilde{\gamma}_{m-1} + \widetilde{e}_i$$

which has essentially approximated $G(\cdot)$ by a m'th order polynomial

20.5 Least Absolute Deviations

We stated that a conventional goal in econometrics is estimation of impact of variation in $\boldsymbol{x}_i$ on the central tendency of y_i. We have discussed projections and conditional means, but these are not the only measures of central tendency. An alternative good measure is the conditional median.

To recall the definition and properties of the median, let y be a continuous random variable. The median $\theta = \text{med}(y)$ is the value such that $\Pr(y \le \theta) = \Pr(y \ge \theta) = 0.5$. Two useful facts about the median are that

$$\theta = \underset{\theta}{\text{argmin}}\, \mathbb{E}\left|y - \theta\right| \tag{20.11}$$

and

$$\mathbb{E}\left(\text{sgn}\left(y - \theta\right)\right) = 0$$

where

$$\text{sgn}\left(u\right) = \begin{cases} 1 & \text{if } u \ge 0 \\ -1 & \text{if } u < 0 \end{cases}$$

is the sign function.

These facts and definitions motivate three estimators of θ. The first definition is the $50th$ empirical quantile. The second is the value which minimizes $\frac{1}{n}\sum_{i=1}^n \left|y_i - \theta\right|$, and the third definition is the solution to the moment equation $\frac{1}{n}\sum_{i=1}^n \text{sgn}\left(y_i - \theta\right)$. These distinctions are illusory, however, as these estimators are indeed identical.

Now let's consider the conditional median of y given a random vector $\boldsymbol{x}$. Let $m(\boldsymbol{x}) = \text{med}\left(y \mid \boldsymbol{x}\right)$ denote the conditional median of y given $\boldsymbol{x}$. The linear median regression model takes the form

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \text{med}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \end{aligned}$$

In this model, the linear function $\text{med}\left(y_i \mid \boldsymbol{x}_i = \boldsymbol{x}\right) = \boldsymbol{x}'\boldsymbol{\beta}$ is the conditional median function, and the substantive assumption is that the median function is linear in $\boldsymbol{x}$.

Conditional analogs of the facts about the median are

- $\Pr(y_i \leq \boldsymbol{x}'\boldsymbol{\beta} \mid \boldsymbol{x}_i = \boldsymbol{x}) = \Pr(y_i > \boldsymbol{x}'\boldsymbol{\beta} \mid \boldsymbol{x}_i = \boldsymbol{x}) = .5$
- $\mathbb{E}\left(\operatorname{sgn}(e_i) \mid \boldsymbol{x}_i\right) = 0$
- $\mathbb{E}\left(\boldsymbol{x}_i \operatorname{sgn}(e_i)\right) = 0$
- $\boldsymbol{\beta} = \min_{\boldsymbol{\beta}} \mathbb{E}\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$

These facts motivate the following estimator. Let

$$LAD(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$$

be the average of absolute deviations. The **least absolute deviations** (LAD) estimator of $\boldsymbol{\beta}$ minimizes this function

$$\widehat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} LAD(\boldsymbol{\beta})$$

Equivalently, it is a solution to the moment condition

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \operatorname{sgn}\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right) = 0. \tag{20.12}$$

The LAD estimator has an asymptotic normal distribution.

Theorem 20.5.1 ***Asymptotic Distribution of LAD Estimator***
When the conditional median is linear in $\boldsymbol{x}$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\right)$$

where

$$V = \frac{1}{4}\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}$$

and $f(e \mid \boldsymbol{x})$ *is the conditional density of* e_i *given* $\boldsymbol{x}_i = \boldsymbol{x}$.

The variance of the asymptotic distribution inversely depends on $f\left(0 \mid \boldsymbol{x}\right)$, the conditional density of the error at its median. When $f\left(0 \mid \boldsymbol{x}\right)$ is large, then there are many innovations near to the median, and this improves estimation of the median. In the special case where the error is independent of $\boldsymbol{x}_i$, then $f\left(0 \mid \boldsymbol{x}\right) = f\left(0\right)$ and the asymptotic variance simplifies

$$\boldsymbol{V} = \frac{\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}}{4f\left(0\right)^2} \tag{20.13}$$

This simplification is similar to the simplification of the asymptotic covariance of the OLS estimator under homoskedasticity.

Computation of standard error for LAD estimates typically is based on equation (20.13). The main difficulty is the estimation of $f(0)$, the height of the error density at its median. This can be done with kernel estimation techniques. See Chapter 22. While a complete proof of Theorem 20.5.1 is advanced, we provide a sketch here for completeness.

Proof of Theorem 20.5.1: Similar to NLLS, LAD is an optimization estimator. Let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}_0$.

The first step is to show that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$. The general nature of the proof is similar to that for the NLLS estimator, and is sketched here. For any fixed $\boldsymbol{\beta}$, by the WLLN, $LAD(\boldsymbol{\beta}) \xrightarrow{p} \mathbb{E}\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$. Furthermore, it can be shown that this convergence is uniform in $\boldsymbol{\beta}$. (Proving uniform convergence is more challenging than for the NLLS criterion since the LAD criterion is not differentiable in $\boldsymbol{\beta}$.) It follows that $\hat{\boldsymbol{\beta}}$, the minimizer of $LAD(\boldsymbol{\beta})$, converges in probability to $\boldsymbol{\beta}_0$, the minimizer of $\mathbb{E}\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$.

Since $\operatorname{sgn}(a) = 1 - 2 \cdot 1(a \leq 0)$, (20.12) is equivalent to $\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}) = 0$, where $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{\beta})$ and $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{x}_i\left(1 - 2 \cdot 1\left(y_i \leq \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)$. Let $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\right)$. We need three preliminary results. First, since $\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta}_0)\right) = 0$ and $\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta}_0)\boldsymbol{g}_i(\boldsymbol{\beta}_0)'\right) = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)$, we can apply the central limit theorem (Theorem 6.8.1) and find that

$$\sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) = n^{-1/2}\sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right).$$

Second using the law of iterated expectations and the chain rule of differentiation,

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}'}\mathbb{E}\boldsymbol{x}_i\left(1 - 2 \cdot 1\left(y_i \leq \boldsymbol{x}_i'\boldsymbol{\beta}\right)\right) \\
&= -2\frac{\partial}{\partial \boldsymbol{\beta}'}\mathbb{E}\left(\boldsymbol{x}_i\mathbb{E}\left(1\left(e_i \leq \boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}_0\right) \mid \boldsymbol{x}_i\right)\right) \\
&= -2\frac{\partial}{\partial \boldsymbol{\beta}'}\mathbb{E}\left(\boldsymbol{x}_i\int_{-\infty}^{\boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}_0} f\left(e \mid \boldsymbol{x}_i\right)de\right) \\
&= -2\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'f\left(\boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{x}_i'\boldsymbol{\beta}_0 \mid \boldsymbol{x}_i\right)\right)
\end{aligned}$$

so

$$\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}(\boldsymbol{\beta}) = -2\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'f\left(0 \mid \boldsymbol{x}_i\right)\right).$$

Third, by a Taylor series expansion and the fact $\boldsymbol{g}(\boldsymbol{\beta}) = 0$

$$\boldsymbol{g}(\widehat{\boldsymbol{\beta}}) \simeq \frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}(\boldsymbol{\beta})\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right).$$

Together

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) &\simeq \left(\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}(\boldsymbol{\beta}_0)\right)^{-1}\sqrt{n}\boldsymbol{g}(\widehat{\boldsymbol{\beta}}) \\
&= \left(-2\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}\sqrt{n}\left(\boldsymbol{g}(\widehat{\boldsymbol{\beta}}) - \overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})\right) \\
&\simeq \frac{1}{2}\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}\sqrt{n}\left(\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) - \boldsymbol{g}(\boldsymbol{\beta}_0)\right) \\
&\xrightarrow{d} \frac{1}{2}\left(\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'f\left(0 \mid \boldsymbol{x}_i\right)\right]\right)^{-1}\mathrm{N}\left(\mathbf{0}, \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right) \\
&= \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\right).
\end{aligned}$$

The third line follows from an asymptotic empirical process argument and the fact that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.

20.6 Quantile Regression

Quantile regression has become quite popular in recent econometric practice. For $\tau \in [0, 1]$ the τ^{th} quantile Q_τ of a random variable with distribution function $F(u)$ is defined as

$$Q_\tau = \inf\left\{u : F(u) \geq \tau\right\}$$

When $F(u)$ is continuous and strictly monotonic, then $F\left(Q_\tau\right) = \tau$, so you can think of the quantile as the inverse of the distribution function. The quantile Q_τ is the value such that τ (percent) of the mass of the distribution is less than Q_τ. The median is the special case $\tau = .5$.

The following alternative representation is useful. If the random variable U has τ^{th} quantile Q_τ, then

$$Q_\tau = \underset{\theta}{\operatorname{argmin}} \mathbb{E}\left(\rho_\tau\left(U - \theta\right)\right). \tag{20.14}$$

where $\rho_\tau\left(q\right)$ is the piecewise linear function

$$\begin{aligned} \rho_\tau\left(q\right) &= \begin{cases} -q\left(1-\tau\right) & q < 0 \\ q\tau & q \geq 0 \end{cases} \\ &= q\left(\tau - 1\left(q < 0\right)\right). \end{aligned} \tag{20.15}$$

This generalizes representation (20.11) for the median to all quantiles.

For the random variables $(y_i, \boldsymbol{x}_i)$ with conditional distribution function $F\left(y \mid \boldsymbol{x}\right)$ the conditional quantile function $q_\tau(\boldsymbol{x})$ is

$$Q_\tau(\boldsymbol{x}) = \inf\left\{y : F\left(y \mid \boldsymbol{x}\right) \geq \tau\right\}.$$

Again, when $F\left(y \mid \boldsymbol{x}\right)$ is continuous and strictly monotonic in y, then $F\left(Q_\tau(\boldsymbol{x}) \mid \boldsymbol{x}\right) = \tau$. For fixed τ, the quantile regression function $q_\tau(\boldsymbol{x})$ describes how the τ^{th} quantile of the conditional distribution varies with the regressors.

As functions of $\boldsymbol{x}$, the quantile regression functions can take any shape. However for computational convenience it is typical to assume that they are (approximately) linear in $\boldsymbol{x}$ (after suitable transformations). This linear specification assumes that $Q_\tau(\boldsymbol{x}) = \boldsymbol{\beta}_\tau'\boldsymbol{x}$ where the coefficients $\boldsymbol{\beta}_\tau$ vary across the quantiles τ. We then have the linear quantile regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}_\tau + e_i$$

where e_i is the error defined to be the difference between y_i and its τ^{th} conditional quantile $\boldsymbol{x}_i'\boldsymbol{\beta}_\tau$. By construction, the τ^{th} conditional quantile of e_i is zero, otherwise its properties are unspecified without further restrictions.

Given the representation (20.14), the quantile regression estimator $\widehat{\boldsymbol{\beta}}_\tau$ for $\boldsymbol{\beta}_\tau$ solves the minimization problem

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, S^\tau(\boldsymbol{\beta})$$

where

$$S^\tau(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)$$

and $\rho_\tau\left(q\right)$ is defined in (20.15).

Since the quantile regression criterion function $S^\tau(\boldsymbol{\beta})$ does not have an algebraic solution, numerical methods are necessary for its minimization. Furthermore, since it has discontinuous derivatives, conventional Newton-type optimization methods are inappropriate. Fortunately, fast linear programming methods have been developed for this problem, and are widely available.

An asymptotic distribution theory for the quantile regression estimator can be derived using similar arguments as those for the LAD estimator in Theorem 20.5.1.

Theorem 20.6.1 ***Asymptotic Distribution of the Quantile Regression Estimator***
When the τ^{th} conditional quantile is linear in $\boldsymbol{x}$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}_\tau\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_\tau\right),$$

where

$$\boldsymbol{V}_\tau = \tau\left(1-\tau\right)\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}$$

and $f\left(e \mid \boldsymbol{x}\right)$ is the conditional density of e_i given $\boldsymbol{x}_i = \boldsymbol{x}$.

In general, the asymptotic variance depends on the conditional density of the quantile regression error. When the error e_i is independent of $\boldsymbol{x}_i$, then $f\left(0 \mid \boldsymbol{x}_i\right) = f\left(0\right)$, the unconditional density of e_i at 0, and we have the simplification

$$\boldsymbol{V}_\tau = \frac{\tau\left(1-\tau\right)}{f\left(0\right)^2}\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}.$$

A recent monograph on the details of quantile regression is Koenker (2005).

Exercises

Exercise 20.1 Suppose that $y_i = g(\boldsymbol{x}_i, \boldsymbol{\theta}) + e_i$ with $\mathbb{E}(e_i \mid \boldsymbol{x}_i) = 0$, $\widehat{\boldsymbol{\theta}}$ is the NLLS estimator, and $\widehat{\boldsymbol{V}}$ is the estimate of var $\left(\widehat{\boldsymbol{\theta}}\right)$. You are interested in the conditional mean function $\mathbb{E}(y_i \mid \boldsymbol{x}_i = \boldsymbol{x}) = g(\boldsymbol{x})$ at some $\boldsymbol{x}$. Find an asymptotic 95% confidence interval for $g(\boldsymbol{x})$.

Exercise 20.2 In Exercise 9.26, you estimated a cost function on a cross-section of electric companies. The equation you estimated was

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + e_i. \tag{20.16}$$

(a) Following Nerlove, add the variable $(\log Q_i)^2$ to the regression. Do so. Assess the merits of this new specification using a hypothesis test. Do you agree with this modification?

(b) Now try a non-linear specification. Consider model (20.16) plus the extra term $\beta_6 z_i$, where

$$z_i = \log Q_i \left(1 + \exp\left(-\left(\log Q_i - \beta_7\right)\right)\right)^{-1}.$$

In addition, impose the restriction $\beta_3 + \beta_4 + \beta_5 = 1$. This model is called a smooth threshold model. For values of $\log Q_i$ much below β_7, the variable $\log Q_i$ has a regression slope of β_2. For values much above β_7, the regression slope is $\beta_2 + \beta_6$, and the model imposes a smooth transition between these regimes. The model is non-linear because of the parameter β_7.

The model works best when β_7 is selected so that several values (in this example, at least 10 to 15) of $\log Q_i$ are both below and above β_7. Examine the data and pick an appropriate range for β_7.

(c) Estimate the model by non-linear least squares. I recommend the concentration method: Pick 10 (or more if you like) values of β_7 in this range. For each value of β_7, calculate z_i and estimate the model by OLS. Record the sum of squared errors, and find the value of β_7 for which the sum of squared errors is minimized.

(d) Calculate standard errors for all the parameters $(\beta_1, ..., \beta_7)$.

Exercise 20.3 Using the CPS data set, return to the linear regression model reported in Table 4.1

(a) Re-estimate the model by least-squares. You do not need to report the estimates, but confirm that you obtain the same results.

(b) Test whether the error variance is different for men and women. Interpret.

(c) Test whether the error variance is different across the race groups (White, Black, American Indian, Asian, Mixed Race). Interpret.

(d) Construct a model for the conditional variance. Estimate such a model, test for general heteroskedasticity and report the results.

(e) Using this model for the conditional variance, re-estimate the model from part (c) using FGLS. Report the results.

(f) Do the OLS and FGLS estimates differ greatly? Note any interesting differences.

(g) Compare the estimated standard errors. Note any interesting differences.

Exercise 20.4 For any predictor $g(\boldsymbol{x}_i)$ for y_i, the mean absolute error (MAE) is

$$\mathbb{E}\left|y_i - g(\boldsymbol{x}_i)\right|.$$

Show that the function $g(\boldsymbol{x})$ which minimizes the MAE is the conditional median $m(\boldsymbol{x}) = \text{med}(y_i \mid \boldsymbol{x}_i)$.

Exercise 20.5 Define

$$g(u) = \tau - 1(u < 0)$$

where $1(\cdot)$ is the indicator function (takes the value 1 if the argument is true, else equals zero). Let θ satisfy $\mathbb{E}(g(y_i - \theta)) = 0$. Is θ a quantile of the distribution of y_i?

Exercise 20.6 Verify equation (20.14)

Exercise 20.7 You are interested in estimating the equation $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$. You believe the regressors are exogenous, but you are uncertain about the properties of the error. You estimate the equation both by least absolute deviations (LAD) and OLS. A colleagye suggests that you should prefer the OLS estimate, because it produces a higher R^2 than the LAD estimate. Is your colleague correct?

Chapter 21

Limited Dependent Variables

y is a **limited dependent variable** if it takes values in a strict subset of $\mathbb{R}$. The most common cases are

- Binary: $y \in \{0, 1\}$
- Multinomial: $y \in \{0, 1, 2, ..., k\}$
- Integer: $y \in \{0, 1, 2, ...\}$
- Censored: $y \in \mathbb{R}^+$

The traditional approach to the estimation of limited dependent variable (LDV) models is parametric maximum likelihood. A parametric model is constructed, allowing the construction of the likelihood function. A more modern approach is semi-parametric, eliminating the dependence on a parametric distributional assumption. We will discuss only the first (parametric) approach, due to time constraints. They still constitute the majority of LDV applications. If, however, you were to write a thesis involving LDV estimation, you would be advised to consider employing a semi-parametric estimation approach.

For the parametric approach, estimation is by MLE. A major practical issue is construction of the likelihood function.

21.1 Binary Choice

The dependent variable $y_i \in \{0, 1\}$. This represents a Yes/No outcome. Given some regressors $\boldsymbol{x}_i$, the goal is to describe $\Pr(y_i = 1 \mid \boldsymbol{x}_i)$, as this is the full conditional distribution.

The linear probability model specifies that

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

As $\Pr(y_i = 1 \mid \boldsymbol{x}_i) = \mathbb{E}(y_i \mid \boldsymbol{x}_i)$, this yields the regression: $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ which can be estimated by OLS. However, the linear probability model does not impose the restriction that $0 \leq \Pr(y_i \mid \boldsymbol{x}_i) \leq 1$. Even so estimation of a linear probability model is a useful starting point for subsequent analysis.

The standard alternative is to use a function of the form

$$\Pr(y_i = 1 \mid \boldsymbol{x}_i) = F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)$$

where $F(\cdot)$ is a known CDF, typically assumed to be symmetric about zero, so that $F(u) = 1 - F(-u)$. The two standard choices for F are

- Logistic: $F(u) = (1 + e^{-u})^{-1}$.

- Normal: $F(u) = \Phi(u)$.

If F is logistic, we call this the **logit** model, and if F is normal, we call this the **probit** model. This model is identical to the latent variable model

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ e_i &\sim F(\cdot) \\ y_i &= \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

For then

$$\begin{aligned} \Pr(y_i = 1 \mid \boldsymbol{x}_i) &= \Pr(y_i^* > 0 \mid \boldsymbol{x}_i) \\ &= \Pr\left(\boldsymbol{x}_i'\boldsymbol{\beta} + e_i > 0 \mid \boldsymbol{x}_i\right) \\ &= \Pr\left(e_i > -\boldsymbol{x}_i'\boldsymbol{\beta} \mid \boldsymbol{x}_i\right) \\ &= 1 - F\left(-\boldsymbol{x}_i'\boldsymbol{\beta}\right) \\ &= F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right). \end{aligned}$$

Estimation is by maximum likelihood. To construct the likelihood, we need the conditional distribution of an individual observation. Recall that if y is Bernoulli, such that $\Pr(y = 1) = p$ and $\Pr(y = 0) = 1 - p$, then we can write the density of y as

$$f(y) = p^y(1-p)^{1-y}, \qquad y = 0, 1.$$

In the Binary choice model, y_i is conditionally Bernoulli with $\Pr(y_i = 1 \mid \boldsymbol{x}_i) = p_i = F(\boldsymbol{x}_i'\boldsymbol{\beta})$. Thus the conditional density is

$$\begin{aligned} f(y_i \mid \boldsymbol{x}_i) &= p_i^{y_i}(1-p_i)^{1-y_i} \\ &= F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)^{y_i}\left(1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{1-y_i}. \end{aligned}$$

Hence the log-likelihood function is

$$\begin{aligned} \log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{x}_i) \\ &= \sum_{i=1}^{n} \log\left(F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)^{y_i}\left(1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^{1-y_i}\right) \\ &= \sum_{i=1}^{n} \left[y_i \log F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right) + (1-y_i)\log(1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right))\right] \\ &= \sum_{y_i=1} \log F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right) + \sum_{y_i=0} \log(1 - F\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)). \end{aligned}$$

The MLE $\widehat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ which maximizes $\log L(\boldsymbol{\beta})$. Standard errors and test statistics are computed by asymptotic approximations. Details of such calculations are left to more advanced courses.

21.2 Count Data

If $y \in \{0, 1, 2, ...\}$, a typical approach is to employ **Poisson regression**. This model specifies that

$$\begin{aligned} \Pr(y_i = k \mid \boldsymbol{x}_i) &= \frac{\exp(-\lambda_i)\lambda_i^k}{k!}, \qquad k = 0, 1, 2, ... \\ \lambda_i &= \exp(\boldsymbol{x}_i'\boldsymbol{\beta}). \end{aligned}$$

The conditional density is the Poisson with parameter λ_i. The functional form for λ_i has been picked to ensure that $\lambda_i > 0$.

The log-likelihood function is

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{x}_i) = \sum_{i=1}^{n} \left(-\exp(\boldsymbol{x}_i'\boldsymbol{\beta}) + y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \log(y_i!) \right).$$

The MLE is the value $\hat{\boldsymbol{\beta}}$ which maximizes $\log L(\boldsymbol{\beta})$.

Since

$$\mathbb{E}(y_i \mid \boldsymbol{x}_i) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

is the conditional mean, this motivates the label Poisson "regression."

Also observe that the model implies that

$$\operatorname{var}(y_i \mid \boldsymbol{x}_i) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}),$$

so the model imposes the restriction that the conditional mean and variance of y_i are the same. This may be considered restrictive. A generalization is the negative binomial.

21.3 Censored Data

The idea of **censoring** is that some data above or below a threshold are mis-reported at the threshold. Thus the model is that there is some latent process y_i^* with unbounded support, but we observe only

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}. \tag{21.1}$$

(This is written for the case of the threshold being zero, any known value can substitute.) The observed data y_i therefore come from a mixed continuous/discrete distribution.

Censored models are typically applied when the data set has a meaningful proportion (say 5% or higher) of data at the boundary of the sample support. The censoring process may be explicit in data collection, or it may be a by-product of economic constraints.

An example of a data collection censoring is top-coding of income. In surveys, incomes above a threshold are typically reported at the threshold.

The first censored regression model was developed by Tobin (1958) to explain consumption of durable goods. Tobin observed that for many households, the consumption level (purchases) in a particular period was zero. He proposed the latent variable model

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ e_i &\overset{iid}{\sim} \mathrm{N}(0, \sigma^2) \end{aligned}$$

with the observed variable y_i generated by the censoring equation (21.1). This model (now called the Tobit) specifies that the latent (or ideal) value of consumption may be negative (the household would prefer to sell than buy). All that is reported is that the household purchased zero units of the good.

The naive approach to estimate $\boldsymbol{\beta}$ is to regress y_i on $\boldsymbol{x}_i$. This does not work because regression estimates $\mathbb{E}(y_i \mid \boldsymbol{x}_i)$, not $\mathbb{E}(y_i^* \mid \boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, and the latter is of interest. Thus OLS will be biased for the parameter of interest $\boldsymbol{\beta}$.

[Note: it is still possible to estimate $\mathbb{E}(y_i \mid \boldsymbol{x}_i)$ by LS techniques. The Tobit framework postulates that this is not inherently interesting, that the parameter of $\boldsymbol{\beta}$ is defined by an alternative statistical structure.]

Consistent estimation will be achieved by the MLE. To construct the likelihood, observe that the probability of being censored is

$$\begin{aligned}\Pr\left(y_i = 0 \mid \boldsymbol{x}_i\right) &= \Pr\left(y_i^* < 0 \mid \boldsymbol{x}_i\right) \\ &= \Pr\left(\boldsymbol{x}_i'\boldsymbol{\beta} + e_i < 0 \mid \boldsymbol{x}_i\right) \\ &= \Pr\left(\frac{e_i}{\sigma} < -\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma} \mid \boldsymbol{x}_i\right) \\ &= \Phi\left(-\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right).\end{aligned}$$

The conditional density function above zero is normal:

$$\sigma^{-1}\phi\left(\frac{y - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right), \qquad y > 0.$$

Therefore, the density function for $y \geq 0$ can be written as

$$f\left(y \mid \boldsymbol{x}_i\right) = \Phi\left(-\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)^{1(y=0)} \left[\sigma^{-1}\phi\left(\frac{z - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right]^{1(y>0)},$$

where $1\left(\cdot\right)$ is the indicator function.

Hence the log-likelihood is a mixture of the probit and the normal:

$$\begin{aligned}\log L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{x}_i) \\ &= \sum_{y_i=0} \log \Phi\left(-\frac{\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right) + \sum_{y_i>0} \log\left[\sigma^{-1}\phi\left(\frac{y_i - \boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right].\end{aligned}$$

The MLE is the value $\widehat{\boldsymbol{\beta}}$ which maximizes $\log L(\boldsymbol{\beta})$.

21.4 Sample Selection

The problem of sample selection arises when the sample is a non-random selection of potential observations. This occurs when the observed data is systematically different from the population of interest. For example, if you ask for volunteers for an experiment, and they wish to extrapolate the effects of the experiment on a general population, you should worry that the people who volunteer may be systematically different from the general population. This has great relevance for the evaluation of anti-poverty and job-training programs, where the goal is to assess the effect of "training" on the general population, not just on the volunteers.

A simple sample selection model can be written as the latent model

$$\begin{aligned}y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_{1i} \\ T_i &= 1\left(\boldsymbol{z}_i'\boldsymbol{\gamma} + e_{0i} > 0\right)\end{aligned}$$

where $1\left(\cdot\right)$ is the indicator function. The dependent variable y_i is observed if (and only if) $T_i = 1$. Else it is unobserved.

For example, y_i could be a wage, which can be observed only if a person is employed. The equation for T_i is an equation specifying the probability that the person is employed.

The model is often completed by specifying that the errors are jointly normal

$$\begin{pmatrix} e_{0i} \\ e_{1i} \end{pmatrix} \sim \mathrm{N}\left(0, \begin{pmatrix} 1 & \rho \\ \rho & \sigma^2 \end{pmatrix}\right).$$

It is presumed that we observe $\{\boldsymbol{x}_i, \boldsymbol{z}_i, T_i\}$ for all observations.

Under the normality assumption,

$$e_{1i} = \rho e_{0i} + v_i,$$

where v_i is independent of $e_{0i} \sim \mathrm{N}(0,1)$. A useful fact about the standard normal distribution is that

$$\mathbb{E}\left(e_{0i} \mid e_{0i} > -x\right) = \lambda(x) = \frac{\phi(x)}{\Phi(x)},$$

and the function $\lambda(x)$ is called the inverse Mills ratio.

The naive estimator of $\boldsymbol{\beta}$ is OLS regression of y_i on $\boldsymbol{x}_i$ for those observations for which y_i is available. The problem is that this is equivalent to conditioning on the event $\{T_i = 1\}$. However,

$$\begin{aligned}
\mathbb{E}\left(e_{1i} \mid T_i = 1, \boldsymbol{z}_i\right) &= \mathbb{E}\left(e_{1i} \mid \{e_{0i} > -\boldsymbol{z}_i'\boldsymbol{\gamma}\}, \boldsymbol{z}_i\right) \\
&= \rho\mathbb{E}\left(e_{0i} \mid \{e_{0i} > -\boldsymbol{z}_i'\boldsymbol{\gamma}\}, \boldsymbol{z}_i\right) + \mathbb{E}\left(v_i \mid \{e_{0i} > -\boldsymbol{z}_i'\boldsymbol{\gamma}\}, \boldsymbol{z}_i\right) \\
&= \rho\lambda\left(\boldsymbol{z}_i'\boldsymbol{\gamma}\right),
\end{aligned}$$

which is non-zero. Thus

$$e_{1i} = \rho\lambda\left(\boldsymbol{z}_i'\boldsymbol{\gamma}\right) + u_i,$$

where

$$\mathbb{E}\left(u_i \mid T_i = 1, \boldsymbol{z}_i\right) = 0.$$

Hence

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \rho\lambda\left(\boldsymbol{z}_i'\boldsymbol{\gamma}\right) + u_i \tag{21.2}$$

is a valid regression equation for the observations for which $T_i = 1$.

Heckman (1979) observed that we could consistently estimate $\boldsymbol{\beta}$ and ρ from this equation, if $\boldsymbol{\gamma}$ were known. It is unknown, but also can be consistently estimated by a Probit model for selection. The "Heckit" estimator is thus calculated as follows

- Estimate $\widehat{\boldsymbol{\gamma}}$ from a Probit, using regressors $\boldsymbol{z}_i$. The binary dependent variable is T_i.
- Estimate $\left(\widehat{\boldsymbol{\beta}}, \widehat{\rho}\right)$ from OLS of y_i on $\boldsymbol{x}_i$ and $\lambda(\boldsymbol{z}_i'\widehat{\boldsymbol{\gamma}})$.
- The OLS standard errors will be incorrect, as this is a two-step estimator. They can be corrected using a more complicated formula. Or, alternatively, by viewing the Probit/OLS estimation equations as a large joint GMM problem.

The Heckit estimator is frequently used to deal with problems of sample selection. However, the estimator is built on the assumption of normality, and the estimator can be quite sensitive to this assumption. Some modern econometric research is exploring how to relax the normality assumption.

The estimator can also work quite poorly if $\lambda\left(\boldsymbol{z}_i'\widehat{\boldsymbol{\gamma}}\right)$ does not have much in-sample variation. This can happen if the Probit equation does not "explain" much about the selection choice. Another potential problem is that if $\boldsymbol{z}_i = \boldsymbol{x}_i$, then $\lambda\left(\boldsymbol{z}_i'\widehat{\boldsymbol{\gamma}}\right)$ can be highly collinear with $\boldsymbol{x}_i$, so the second step OLS estimator will not be able to precisely estimate $\boldsymbol{\beta}$. Based this observation, it is typically recommended to find a valid exclusion restriction: a variable should be in $\boldsymbol{z}_i$ which is not in $\boldsymbol{x}_i$. If this is valid, it will ensure that $\lambda\left(\boldsymbol{z}_i'\widehat{\boldsymbol{\gamma}}\right)$ is not collinear with $\boldsymbol{x}_i$, and hence improve the second stage estimator's precision.

Exercises

Exercise 21.1 Your model is

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \end{aligned}$$

However, y_i^* is not observed. Instead only a capped version is reported. That is, the dataset contains the variable

$$y_i = \begin{cases} y_i^* & \text{if} \quad y_i^* \leq \tau \\ \\ \tau & \text{if} \quad y_i^* > \tau \end{cases}$$

Suppose you regress y_i on x_i using OLS. Is OLS consistent for $\boldsymbol{\beta}$? Describe the nature of the effect of the mis-measured observation on the OLS estimate.

Exercise 21.2 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) &= 0 \end{aligned}$$

Let $\widehat{\boldsymbol{\beta}}$ denote the OLS estimator for $\boldsymbol{\beta}$ based on an available sample.

(a) Suppose that the i^{th} observation is in the sample only if $x_{1i} > 0$, where x_{1i} is an element of x_i. Assume $\Pr\left(x_{1i} < 0\right) > 0$.

i Is $\widehat{\boldsymbol{\beta}}$ consistent for $\widehat{\boldsymbol{\beta}}$?

ii If not, can you obtain an expression for its probability limit?
(For this, you may assume that e_i is independent of $\boldsymbol{x}_i$ and $N(0, \sigma^2)$.)

(b) Suppose that the i^{th} observation is in the sample only if $y_i > 0$.

i Is $\widehat{\boldsymbol{\beta}}$ consistent for $\widehat{\boldsymbol{\beta}}$?

ii If not, can you obtain an expression for its probability limit?
(For this, you may assume that e_i is independent of $\boldsymbol{x}_i$ and $\mathrm{N}(0, \sigma^2)$.)

Exercise 21.3 The Tobit model is

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ e_i &\sim \mathrm{N}\left(0, \sigma^2\right) \\ y_i &= y_i^* 1\left(y_i^* \geq 0\right) \end{aligned}$$

where $1(\cdot)$ is the indicator function.

(a) Find $\mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right)$.

Note: You may use the fact that since $e_i \sim N\left(0, \sigma^2\right)$,

$$\mathbb{E}\left(e_i 1\left(e_i \geq -u\right)\right) = \sigma\lambda(u/\sigma) = \sigma\phi(u/\sigma)/\Phi(u/\sigma).$$

(b) Use the result from part (a) to suggest a NLLS estimator for the parameter β given a sample $\{y_i, \boldsymbol{x}_i\}$.

Exercise 21.4 A latent variable y_i^* is generated by

$$y_i^* = x_i\beta + e_i$$

The distribution of e_i, conditional on x_i, is $\mathrm{N}(0, \sigma_i^2)$, where $\sigma_i^2 = \gamma_0 + x_i^2\gamma_1$ with $\gamma_0 > 0$ and $\gamma_1 > 0$. The binary variable y_i equals 1 if $y_i^* \geq 0$, else $y_i = 0$. Find the log-likelihood function for the conditional distribution of y_i given x_i (the parameters are $\beta, \gamma_0, \gamma_1$).

Chapter 22

Nonparametric Density Estimation

22.1 Kernel Density Estimation

Let X be a random variable with continuous distribution $F(x)$ and density $f(x) = \frac{d}{dx}F(x)$. The goal is to estimate $f(x)$ from a random sample $(X_1, ..., X_n\}$ While $F(x)$ can be estimated by the EDF $\widehat{F}(x) = n^{-1}\sum_{i=1}^{n} 1(X_i \leq x)$, we cannot define $\frac{d}{dx}\widehat{F}(x)$ since $\widehat{F}(x)$ is a step function. The standard **nonparametric** method to estimate $f(x)$ is based on **smoothing** using a kernel.

While we are typically interested in estimating the entire function $f(x)$, we can simply focus on the problem where x is a specific fixed number, and then see how the method generalizes to estimating the entire function.

The most common methods to estimate the density $f(x)$ is by kernel methods, which are similar to the nonparametric methods introduced in Section 17. As for kernel regression, density estimation uses kernel functions $k(u)$, which are density functions symmetric about zero. See Section 17 for a discussion of kernel functions.

The kernel functions are used to smooth the data. The amount of smoothing is controlled by the **bandwidth** $h > 0$. Define the rescaled kernel function

$$k_h(u) = \frac{1}{h}k\left(\frac{u}{h}\right).$$

The kernel density estimator of $f(x)$ is

$$\widehat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(X_i - x).$$

This estimator is the average of a set of weights. If a large number of the observations X_i are near x, then the weights are relatively large and $\hat{f}(x)$ is larger. Conversely, if only a few X_i are near x, then the weights are small and $\widehat{f}(x)$ is small. The bandwidth h controls the meaning of "near".

Interestingly, if $k(u)$ is a second-order kernel then $\widehat{f}(x)$ is a valid density. That is, $\widehat{f}(x) \geq 0$ for all x, and

$$\begin{aligned}\int_{-\infty}^{\infty} \widehat{f}(x)dx &= \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1}^{n} k_h(X_i - x)\,dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} k_h(X_i - x)\,dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} k(u)\,du = 1\end{aligned}$$

where the second-to-last equality makes the change-of-variables $u = (X_i - x)/h$.

We can also calculate the moments of the density $\widehat{f}(x)$. The mean is

$$\begin{aligned}\int_{-\infty}^{\infty} x\widehat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} xk_h\left(X_i - x\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(X_i - uh\right)k\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i\int_{-\infty}^{\infty} k\left(u\right)du + \frac{1}{n}\sum_{i=1}^{n} h\int_{-\infty}^{\infty} uk\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i\end{aligned}$$

the sample mean of the X_i, where the second-to-last equality used the change-of-variables $u = (X_i - x)/h$ which has Jacobian h.

The second moment of the estimated density is

$$\begin{aligned}\int_{-\infty}^{\infty} x^2\widehat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x^2k_h\left(X_i - x\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(X_i - uh\right)^2k\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{i=1}^{n} X_ih\int_{-\infty}^{\infty} uk(u)du + \frac{1}{n}\sum_{i=1}^{n} h^2\int_{-\infty}^{\infty} u^2k\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\sigma_k^2\end{aligned}$$

where

$$\sigma_K^2 = \int_{-\infty}^{\infty} u^2k\left(u\right)du$$

is the variance of the kernel (see Section 17). It follows that the variance of the density $\widehat{f}(x)$ is

$$\begin{aligned}\int_{-\infty}^{\infty} x^2\widehat{f}(x)dx - \left(\int_{-\infty}^{\infty} x\widehat{f}(x)dx\right)^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\sigma_k^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\ &= \widehat{\sigma}^2 + h^2\sigma_k^2\end{aligned}$$

Thus the variance of the estimated density is inflated by the factor $h^2\sigma_K^2$ relative to the sample moment.

22.2 Asymptotic MSE for Kernel Estimates

For fixed x and bandwidth h observe that

$$\begin{aligned}\mathbb{E}k_h\left(X - x\right) &= \int_{-\infty}^{\infty} k_h\left(z - x\right)f(z)dz \\ &= \int_{-\infty}^{\infty} k_h\left(uh\right)f(x + hu)hdu \\ &= \int_{-\infty}^{\infty} k\left(u\right)f(x + hu)du\end{aligned}$$

The second equality uses the change-of variables $u = (z - x)/h$. The last expression shows that the expected value is an average of $f(z)$ locally about x.

This integral (typically) is not analytically solvable, so we approximate it using a second order Taylor expansion of $f(x + hu)$ in the argument hu about $hu = 0$, which is valid as $h \to 0$. Thus

$$f(x + hu) \simeq f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2$$

and therefore

$$\begin{aligned}
\mathbb{E}k_h(X - x) &\simeq \int_{-\infty}^{\infty} k(u)\left(f(x) + f'(x)hu + \frac{1}{2}f''(x)h^2u^2\right)du \\
&= f(x)\int_{-\infty}^{\infty} k(u)\,du + f'(x)h\int_{-\infty}^{\infty} k(u)\,udu \\
&+ \frac{1}{2}f''(x)h^2\int_{-\infty}^{\infty} k(u)\,u^2du \\
&= f(x) + \frac{1}{2}f''(x)h^2\sigma_k^2.
\end{aligned}$$

The bias of $\widehat{f}(x)$ is then

$$Bias(x) = \mathbb{E}\left(\widehat{f}(x)\right) - f(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(k_h(X_i - x)\right) - f(x) = \frac{1}{2}f''(x)h^2\sigma_k^2.$$

We see that the bias of $\widehat{f}(x)$ at x depends on the second derivative $f''(x)$. The sharper the derivative, the greater the bias. Intuitively, the estimator $\widehat{f}(x)$ smooths data local to $X_i = x$, so is estimating a smoothed version of $f(x)$. The bias results from this smoothing, and is larger the greater the curvature in $f(x)$.

We now examine the variance of $\widehat{f}(x)$. Since it is an average of iid random variables, using first-order Taylor approximations and the fact that n^{-1} is of smaller order than $(nh)^{-1}$

$$\begin{aligned}
\operatorname{var}\left(\widehat{f}(x)\right) &= \frac{1}{n}\operatorname{var}\left(k_h(X_i - x)\right) \\
&= \frac{1}{n}\mathbb{E}\left(k_h(X_i - x)^2\right) - \frac{1}{n}\left(\mathbb{E}\left(k_h(X_i - x)\right)\right)^2 \\
&\simeq \frac{1}{nh^2}\int_{-\infty}^{\infty} k\left(\frac{z - x}{h}\right)^2 f(z)dz - \frac{1}{n}f(x)^2 \\
&= \frac{1}{nh}\int_{-\infty}^{\infty} k(u)^2 f(x + hu)\,du \\
&\simeq \frac{f(x)}{nh}\int_{-\infty}^{\infty} k(u)^2\,du \\
&= \frac{f(x)R_k}{nh}.
\end{aligned}$$

where $R_k = \int_{-\infty}^{\infty} k(u)^2\,du$ is called the **roughness** of k (see Section 17).

Together, the asymptotic mean-squared error (AMSE) for fixed x is the sum of the approximate squared bias and approximate variance

$$AMSE_h(x) = \frac{1}{4}f''(x)^2h^4\sigma_k^4 + \frac{f(x)R_k}{nh}.$$

A global measure of precision is the asymptotic mean integrated squared error (AMISE)

$$AMISE_h = \int AMSE_h(x)dx = \frac{h^4\sigma_k^4R(f'')}{4} + \frac{R_k}{nh}. \tag{22.1}$$

where $R(f'') = \int (f''(x))^2 dx$ is the roughness of f''. Notice that the first term (the squared bias) is increasing in h and the second term (the variance) is decreasing in nh. Thus for the AMISE to decline with n, we need $h \to 0$ but $nh \to \infty$. That is, h must tend to zero, but at a slower rate than n^{-1}.

Equation (22.1) is an asymptotic approximation to the MSE. We define the asymptotically optimal bandwidth h_0 as the value which minimizes this approximate MSE. That is,

$$h_0 = \underset{h}{\operatorname{argmin}}\, AMISE_h$$

It can be found by solving the first order condition

$$\frac{d}{dh} AMISE_h = h^3 \sigma_K^4 R(f'') - \frac{R_k}{nh^2} = 0$$

yielding

$$h_0 = \left(\frac{R_k}{\sigma_k^4 R(f'')} \right)^{1/5} n^{-1/5}. \tag{22.2}$$

This solution takes the form $h_0 = cn^{-1/5}$ where c is a function of k and f, but not of n. We thus say that the optimal bandwidth is of order $O(n^{-1/5})$. Note that this h declines to zero, but at a very slow rate.

In practice, how should the bandwidth be selected? This is a difficult problem, and there is a large literature on the subject. The asymptotically optimal choice given in (22.2) depends on R_k, σ_k^2, and $R(f'')$. The first two are determined by the kernel function and are given in Section 17.

An obvious difficulty is that $R(f'')$ is unknown. A classic simple solution proposed by Silverman (1986) has come to be known as the **reference bandwidth** or **Silverman's Rule-of-Thumb**. It uses formula (22.2) but replaces $R(f'')$ with $\widehat{\sigma}^{-5} R(\phi'')$, where ϕ is the N(0, 1) distribution and $\widehat{\sigma}^2$ is an estimate of $\sigma^2 = \operatorname{var}(X)$. This choice for h gives an optimal rule when $f(x)$ is normal, and gives a nearly optimal rule when $f(x)$ is close to normal. The downside is that if the density is very far from normal, the rule-of-thumb h can be quite inefficient. We can calculate that $R(\phi'') = 3/\left(8\sqrt{\pi}\right)$. Together with the above table, we find the reference rules for the three kernel functions introduced earlier.

Gaussian Kernel: $h_{rule} = 1.06\widehat{\sigma} n^{-1/5}$

Epanechnikov Kernel: $h_{rule} = 2.34\widehat{\sigma} n^{-1/5}$

Biweight (Quartic) Kernel: $h_{rule} = 2.78\widehat{\sigma} n^{-1/5}$

Unless you delve more deeply into kernel estimation methods the rule-of-thumb bandwidth is a good practical bandwidth choice, perhaps adjusted by visual inspection of the resulting estimate $\widehat{f}(x)$. There are other approaches, but implementation can be delicate. I now discuss some of these choices. The **plug-in** approach is to estimate $R(f'')$ in a first step, and then plug this estimate into the formula (22.2). This is more treacherous than may first appear, as the optimal h for estimation of the roughness $R(f'')$ is quite different than the optimal h for estimation of $f(x)$. However, there are modern versions of this estimator work well, in particular the iterative method of Sheather and Jones (1991). Another popular choice for selection of h is **cross-validation**. This works by constructing an estimate of the MISE using leave-one-out estimators. There are some desirable properties of cross-validation bandwidths, but they are also known to converge very slowly to the optimal values. They are also quite ill-behaved when the data has some discretization (as is common in economics), in which case the cross-validation rule can sometimes select very small bandwidths leading to dramatically undersmoothed estimates.

Appendix A

Matrix Algebra

A.1 Notation

A **scalar** a is a single number.

A **vector** $\boldsymbol{a}$ is a $k \times 1$ list of numbers, typically arranged in a column. We write this as

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix}$$

Equivalently, a vector $\boldsymbol{a}$ is an element of Euclidean k space, written as $\boldsymbol{a} \in \mathbb{R}^k$. If $k = 1$ then $\boldsymbol{a}$ is a scalar.

A **matrix** $\boldsymbol{A}$ is a $k \times r$ rectangular array of numbers, written as

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & & \vdots \\ a_{k1} & a_{k2} & \cdots & a_{kr} \end{bmatrix}$$

By convention a_{ij} refers to the element in the i^{th} row and j^{th} column of $\boldsymbol{A}$. If $r = 1$ then $\boldsymbol{A}$ is a column vector. If $k = 1$ then $\boldsymbol{A}$ is a row vector. If $r = k = 1$, then $\boldsymbol{A}$ is a scalar.

A standard convention (which we will follow in this text whenever possible) is to denote scalars by lower-case italics (a), vectors by lower-case bold italics ($\boldsymbol{a}$), and matrices by upper-case bold italics ($\boldsymbol{A}$). Sometimes a matrix $\boldsymbol{A}$ is denoted by the symbol (a_{ij}).

A matrix can be written as a set of column vectors or as a set of row vectors. That is,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_r \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{bmatrix}$$

where

$$\boldsymbol{a}_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ki} \end{bmatrix}$$

are column vectors and

$$\boldsymbol{\alpha}_j = \begin{bmatrix} a_{j1} & a_{j2} & \cdots & a_{jr} \end{bmatrix}$$

are row vectors.

The **transpose** of a matrix $\boldsymbol{A}$, denoted $\boldsymbol{A}'$, $\boldsymbol{A}^\top$, or $\boldsymbol{A}^t$, is obtained by flipping the matrix on its diagonal. (In most of the econometrics literature, and this textbook, we use $\boldsymbol{A}'$, but in the mathematics literature $\boldsymbol{A}^\top$ is the convention.) Thus

$$\boldsymbol{A}' = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & & \vdots \\ a_{1r} & a_{2r} & \cdots & a_{kr} \end{bmatrix}$$

Alternatively, letting $\boldsymbol{B} = \boldsymbol{A}'$, then $b_{ij} = a_{ji}$. Note that if $\boldsymbol{A}$ is $k \times r$, then $\boldsymbol{A}'$ is $r \times k$. If $\boldsymbol{a}$ is a $k \times 1$ vector, then $\boldsymbol{a}'$ is a $1 \times k$ row vector.

A matrix is **square** if $k = r$. A square matrix is **symmetric** if $\boldsymbol{A} = \boldsymbol{A}'$, which requires $a_{ij} = a_{ji}$. A square matrix is **diagonal** if the off-diagonal elements are all zero, so that $a_{ij} = 0$ if $i \neq j$. A square matrix is **upper (lower) diagonal** if all elements below (above) the diagonal equal zero.

An important diagonal matrix is the **identity matrix**, which has ones on the diagonal. The $k \times k$ identity matrix is denoted as

$$\boldsymbol{I}_k = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}.$$

A **partitioned matrix** takes the form

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} & \cdots & \boldsymbol{A}_{1r} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} & \cdots & \boldsymbol{A}_{2r} \\ \vdots & \vdots & & \vdots \\ \boldsymbol{A}_{k1} & \boldsymbol{A}_{k2} & \cdots & \boldsymbol{A}_{kr} \end{bmatrix}$$

where the A_{ij} denote matrices, vectors and/or scalars.

A.2 Complex Matrices*

Scalars, vectors and matrices may contain real or complex numbers as entries. (However, most econometric applications exclusively use real matrices.) If all elements of a vector $\boldsymbol{x}$ are real we say that $\boldsymbol{x}$ is a real vector, and similarly for matrices.

Recall that a complex number can be written as $x = a + b\mathrm{i}$ where where $\mathrm{i} = \sqrt{-1}$ and a and b are real numbers. Similarly a vector with complex elements can be written as $\boldsymbol{x} = \boldsymbol{a} + \boldsymbol{b}\mathrm{i}$ where $\boldsymbol{a}$ and $\boldsymbol{b}$ are real vectors, and a matrix with complex elements can be written as $\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{B}\mathrm{i}$ where $\boldsymbol{A}$ and $\boldsymbol{B}$ are real matrices.

Recall that the complex conjugate of $x = a + b\mathrm{i}$ is $x^* = a - b\mathrm{i}$. For matrices, the analogous concept is the conjugate transpose. The conjugate transpose of $\boldsymbol{X} = \boldsymbol{A} + \boldsymbol{B}\mathrm{i}$ is $\boldsymbol{X}^* = \boldsymbol{A}' - \boldsymbol{B}'\mathrm{i}$. It is obtained by taking the transpose and taking the complex conjugate of each element.

A.3 Matrix Addition

If the matrices $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{ij})$ are of the same order, we define the sum

$$\boldsymbol{A} + \boldsymbol{B} = (a_{ij} + b_{ij}).$$

Matrix addition follows the commutative and associative laws:

$$\boldsymbol{A} + \boldsymbol{B} = \boldsymbol{B} + \boldsymbol{A}$$
$$\boldsymbol{A} + (\boldsymbol{B} + \boldsymbol{C}) = (\boldsymbol{A} + \boldsymbol{B}) + \boldsymbol{C}.$$

A.4 Matrix Multiplication

If $\boldsymbol{A}$ is $k \times r$ and c is real, we define their product as

$$\boldsymbol{A}c = c\boldsymbol{A} = (a_{ij}c).$$

If $\boldsymbol{a}$ and $\boldsymbol{b}$ are both $k \times 1$, then their inner product is

$$\boldsymbol{a}'\boldsymbol{b} = a_1b_1 + a_2b_2 + \cdots + a_kb_k = \sum_{j=1}^{k} a_jb_j.$$

Note that $\boldsymbol{a}'\boldsymbol{b} = \boldsymbol{b}'\boldsymbol{a}$. We say that two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ are **orthogonal** if $\boldsymbol{a}'\boldsymbol{b} = 0$.

If $\boldsymbol{A}$ is $k \times r$ and $\boldsymbol{B}$ is $r \times s$, so that the number of columns of $\boldsymbol{A}$ equals the number of rows of $\boldsymbol{B}$, we say that $\boldsymbol{A}$ and $\boldsymbol{B}$ are **conformable**. In this event the matrix product $\boldsymbol{AB}$ is defined. Writing $\boldsymbol{A}$ as a set of row vectors and $\boldsymbol{B}$ as a set of column vectors (each of length r), then the matrix product is defined as

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \\ \vdots \\ \boldsymbol{a}_k' \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_s \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1'\boldsymbol{b}_s \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2'\boldsymbol{b}_s \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_k'\boldsymbol{b}_1 & \boldsymbol{a}_k'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_k'\boldsymbol{b}_s \end{bmatrix}.$$

Matrix multiplication is not commutative: in general $\boldsymbol{AB} \neq \boldsymbol{BA}$**.** However, it is associative and distributive:

$$\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$$
$$\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}.$$

An alternative way to write the matrix product is to use matrix partitions. For example,

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11} + \boldsymbol{A}_{12}\boldsymbol{B}_{21} & \boldsymbol{A}_{11}\boldsymbol{B}_{12} + \boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11} + \boldsymbol{A}_{22}\boldsymbol{B}_{21} & \boldsymbol{A}_{21}\boldsymbol{B}_{12} + \boldsymbol{A}_{22}\boldsymbol{B}_{22} \end{bmatrix}.$$

As another example,

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_r \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \vdots \\ \boldsymbol{B}_r \end{bmatrix}$$
$$= \boldsymbol{A}_1\boldsymbol{B}_1 + \boldsymbol{A}_2\boldsymbol{B}_2 + \cdots + \boldsymbol{A}_r\boldsymbol{B}_r$$
$$= \sum_{j=1}^{r} \boldsymbol{A}_j\boldsymbol{B}_j.$$

An important property of the identity matrix is that if $\boldsymbol{A}$ is $k \times r$, then $\boldsymbol{A}\boldsymbol{I}_r = \boldsymbol{A}$ and $\boldsymbol{I}_k\boldsymbol{A} = \boldsymbol{A}$.

We say two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are **orthogonal** if $\boldsymbol{A}'\boldsymbol{B} = \boldsymbol{0}$. This means that all columns of $\boldsymbol{A}$ are orthogonal with all columns of $\boldsymbol{B}$.

The $k \times r$ matrix $\boldsymbol{H}$, $r \leq k$, is called **orthonormal** if $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_r$. This means that the columns of $\boldsymbol{H}$ are mutually orthogonal, and each column is normalized to have unit length.

A.5 Trace

The **trace** of a $k \times k$ square matrix $\boldsymbol{A}$ is the sum of its diagonal elements

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} a_{ii}.$$

Some straightforward properties for square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and real c are

$$\begin{aligned} \operatorname{tr}(c\boldsymbol{A}) &= c\operatorname{tr}(\boldsymbol{A}) \\ \operatorname{tr}\left(\boldsymbol{A}'\right) &= \operatorname{tr}(\boldsymbol{A}) \\ \operatorname{tr}(\boldsymbol{A}+\boldsymbol{B}) &= \operatorname{tr}(\boldsymbol{A}) + \operatorname{tr}(\boldsymbol{B}) \\ \operatorname{tr}(\boldsymbol{I}_k) &= k. \end{aligned}$$

Also, for $k \times r$ $\boldsymbol{A}$ and $r \times k$ $\boldsymbol{B}$ we have

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}). \tag{A.1}$$

Indeed,

$$\begin{aligned} \operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) &= \operatorname{tr}\begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1'\boldsymbol{b}_k \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2'\boldsymbol{b}_k \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_k'\boldsymbol{b}_1 & \boldsymbol{a}_k'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_k'\boldsymbol{b}_k \end{bmatrix} \\ &= \sum_{i=1}^{k} \boldsymbol{a}_i'\boldsymbol{b}_i \\ &= \sum_{i=1}^{k} \boldsymbol{b}_i'\boldsymbol{a}_i \\ &= \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}). \end{aligned}$$

A.6 Rank and Inverse

The rank of the $k \times r$ matrix ($r \leq k$)

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_r \end{bmatrix}$$

is the number of linearly independent columns $\boldsymbol{a}_j$, and is written as $\operatorname{rank}(\boldsymbol{A})$. We say that $\boldsymbol{A}$ has full rank if $\operatorname{rank}(\boldsymbol{A}) = r$.

A square $k \times k$ matrix $\boldsymbol{A}$ is said to be **nonsingular** if it is has full rank, e.g. $\operatorname{rank}(\boldsymbol{A}) = k$. This means that there is no $k \times 1$ $\boldsymbol{c} \neq \boldsymbol{0}$ such that $\boldsymbol{A}\boldsymbol{c} = \boldsymbol{0}$.

If a square $k \times k$ matrix $\boldsymbol{A}$ is nonsingular then there exists a unique matrix $k \times k$ matrix $\boldsymbol{A}^{-1}$ called the **inverse** of $\boldsymbol{A}$ which satisfies

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k.$$

For non-singular $\boldsymbol{A}$ and $\boldsymbol{C}$, some important properties include

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{A}^{-1} &= \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k \\
\left(\boldsymbol{A}^{-1}\right)' &= \left(\boldsymbol{A}'\right)^{-1} \\
(\boldsymbol{A}\boldsymbol{C})^{-1} &= \boldsymbol{C}^{-1}\boldsymbol{A}^{-1} \\
(\boldsymbol{A}+\boldsymbol{C})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{C}^{-1} \\
\boldsymbol{A}^{-1}-(\boldsymbol{A}+\boldsymbol{C})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{A}^{-1}.
\end{aligned}$$

If a $k \times k$ matrix $\boldsymbol{H}$ is orthonormal (so that $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_k$), then $\boldsymbol{H}$ is nonsingular and $\boldsymbol{H}^{-1} = \boldsymbol{H}'$. Furthermore, $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{I}_k$ and $\boldsymbol{H}'^{-1} = \boldsymbol{H}$.

Another useful result for non-singular $\boldsymbol{A}$ is known as the **Woodbury matrix identity**

$$(\boldsymbol{A}+\boldsymbol{B}\boldsymbol{C}\boldsymbol{D})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{C}\left(\boldsymbol{C}+\boldsymbol{C}\boldsymbol{D}\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{C}\right)^{-1}\boldsymbol{C}\boldsymbol{D}\boldsymbol{A}^{-1}. \tag{A.2}$$

In particular, for $\boldsymbol{C} = -1$, $\boldsymbol{B} = \boldsymbol{b}$ and $\boldsymbol{D} = \boldsymbol{b}'$ for vector $\boldsymbol{b}$ we find what is known as the **Sherman–Morrison formula**

$$\left(\boldsymbol{A}-\boldsymbol{b}\boldsymbol{b}'\right)^{-1} = \boldsymbol{A}^{-1} + \left(1-\boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1}\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}'\boldsymbol{A}^{-1}. \tag{A.3}$$

The following fact about inverting partitioned matrices is quite useful.

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix}^{-1} \overset{def}{=} \begin{bmatrix} \boldsymbol{A}^{11} & \boldsymbol{A}^{12} \\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11\cdot 2}^{-1} & -\boldsymbol{A}_{11\cdot 2}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\ -\boldsymbol{A}_{22\cdot 1}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{A}_{22\cdot 1}^{-1} \end{bmatrix} \tag{A.4}$$

where $\boldsymbol{A}_{11\cdot 2} = \boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}$ and $\boldsymbol{A}_{22\cdot 1} = \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}$. There are alternative algebraic representations for the components. For example, using the Woodbury matrix identity you can show the following alternative expressions

$$\begin{aligned}
\boldsymbol{A}^{11} &= \boldsymbol{A}_{11}^{-1} + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22\cdot 1}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} \\
\boldsymbol{A}^{22} &= \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11\cdot 2}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1} \\
\boldsymbol{A}^{12} &= -\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22\cdot 1}^{-1} \\
\boldsymbol{A}^{21} &= -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11\cdot 2}^{-1}
\end{aligned}$$

Even if a matrix $\boldsymbol{A}$ does not possess an inverse, we can still define the **Moore-Penrose generalized inverse** $\boldsymbol{A}^-$ as the matrix which satisfies

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} &= \boldsymbol{A} \\
\boldsymbol{A}^-\boldsymbol{A}\boldsymbol{A}^- &= \boldsymbol{A}^- \\
\boldsymbol{A}\boldsymbol{A}^- &\text{ is symmetric} \\
\boldsymbol{A}^-\boldsymbol{A} &\text{ is symmetric}
\end{aligned}$$

For any matrix $\boldsymbol{A}$, the Moore-Penrose generalized inverse $\boldsymbol{A}^-$ exists and is unique.

For example, if

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

and $\boldsymbol{A}_{11}^{-1}$ exists then

$$\boldsymbol{A}^- = \begin{bmatrix} \boldsymbol{A}_{11}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}.$$

A.7 Determinant

The **determinant** is a measure of the volume of a square matrix. It is written as $\det \boldsymbol{A}$ or $|\boldsymbol{A}|$.

While the determinant is widely used, its precise definition is rarely needed. However, we present the definition here for completeness. Let $\boldsymbol{A} = (a_{ij})$ be a $k \times k$ matrix . Let $\pi = (j_1, ..., j_k)$ denote a permutation of $(1, ..., k)$. There are $k!$ such permutations. There is a unique count of the number of inversions of the indices of such permutations (relative to the natural order $(1, ..., k)$, and let $\varepsilon_\pi = +1$ if this count is even and $\varepsilon_\pi = -1$ if the count is odd. Then the determinant of $\boldsymbol{A}$ is defined as

$$\det \boldsymbol{A} = \sum_{\pi} \varepsilon_\pi a_{1j_1} a_{2j_2} \cdots a_{kj_k}.$$

For example, if $\boldsymbol{A}$ is 2×2, then the two permutations of $(1, 2)$ are $(1, 2)$ and $(2, 1)$, for which $\varepsilon_{(1,2)} = 1$ and $\varepsilon_{(2,1)} = -1$. Thus

$$\begin{aligned} \det \boldsymbol{A} &= \varepsilon_{(1,2)} a_{11} a_{22} + \varepsilon_{(2,1)} a_{21} a_{12} \\ &= a_{11} a_{22} - a_{12} a_{21}. \end{aligned}$$

For a square matrix $\boldsymbol{A}$, the **minor** M_{ij} of the ij^{th} element a_{ij} is the determinant of the matrix obtained by removing the i^{th} row and j^{th} column of $\boldsymbol{A}$. The **cofactor** of the ij^{th} element is $C_{ij} = (-1)^{i+j} M_{ij}$. An important representation known as Laplace's expansion relates the determinant of $\boldsymbol{A}$ to its cofactors:

$$\det \boldsymbol{A} = \sum_{j=1}^{k} a_{ij} C_{ij}.$$

This holds for all $i = 1, , .., k$. This is often presented as a method for computation of a determinant.

Theorem A.7.1 *Properties of the determinant*

1. $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$
2. $\det(c\boldsymbol{A}) = c^k \det \boldsymbol{A}$
3. $\det(\boldsymbol{AB}) = \det(\boldsymbol{BA}) = (\det \boldsymbol{A})(\det \boldsymbol{B})$
4. $\det\left(\boldsymbol{A}^{-1}\right) = (\det \boldsymbol{A})^{-1}$
5. $\det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = (\det \boldsymbol{D}) \det\left(\boldsymbol{A} - \boldsymbol{BD}^{-1}\boldsymbol{C}\right)$ if $\det \boldsymbol{D} \neq 0$
6. $\det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{D} \end{bmatrix} = \det(\boldsymbol{A})(\det \boldsymbol{D})$ and $\det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \det(\boldsymbol{A})(\det \boldsymbol{D})$
7. If $\boldsymbol{A}$ is $p \times q$ and $\boldsymbol{B}$ is $q \times p$ then $\det(\boldsymbol{I}_p + \boldsymbol{AB}) = \det(\boldsymbol{I}_q + \boldsymbol{BA})$
8. If $\boldsymbol{A}$ and $\boldsymbol{D}$ are invertible then $\det\left(\boldsymbol{A} - \boldsymbol{BD}^{-1}\boldsymbol{C}\right) = \dfrac{\det(\boldsymbol{A})}{\det(\boldsymbol{D})} \det\left(\boldsymbol{D} - \boldsymbol{CA}^{-1}\boldsymbol{B}\right)$
9. $\det \boldsymbol{A} \neq 0$ if and only if $\boldsymbol{A}$ is nonsingular
10. If $\boldsymbol{A}$ is triangular (upper or lower), then $\det \boldsymbol{A} = \prod_{i=1}^{k} a_{ii}$
11. If $\boldsymbol{A}$ is orthonormal, then $\det \boldsymbol{A} = \pm 1$
12. $\boldsymbol{A}^{-1} = (\det \boldsymbol{A})^{-1} \boldsymbol{C}$ where $\boldsymbol{C} = (C_{ij})$ is the matrix of cofactors

A.8 Eigenvalues

The **characteristic equation** of a $k \times k$ square matrix $\boldsymbol{A}$ is

$$\det(\lambda \boldsymbol{I}_k - \boldsymbol{A}) = 0.$$

The left side is a polynomial of degree k in λ so it has exactly k roots, which are not necessarily distinct and may be real or complex. They are called the **latent roots**, **characteristic roots**, or **eigenvalues** of $\boldsymbol{A}$. If λ is an eigenvalue of $\boldsymbol{A}$, then $\lambda \boldsymbol{I}_k - \boldsymbol{A}$ is singular so there exists a non-zero vector $\boldsymbol{h}$ such that $(\lambda \boldsymbol{I}_k - \boldsymbol{A})\boldsymbol{h} = \boldsymbol{0}$ or

$$\boldsymbol{A}\boldsymbol{h} = \boldsymbol{h}\lambda.$$

The vector $\boldsymbol{h}$ is called a **latent vector**, **characteristic vector**, or **eigenvector** of $\boldsymbol{A}$ corresponding to λ. They are typically normalized so that $\boldsymbol{h}'\boldsymbol{h} = 1$ and thus $\lambda = \boldsymbol{h}'\boldsymbol{A}\boldsymbol{h}$.

Set $\boldsymbol{H} = [\boldsymbol{h}_1 \cdots \boldsymbol{h}_k]$ and $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_1, ..., \lambda_k\}$. A matrix expression is

$$\boldsymbol{A}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{\Lambda}$$

We now state some useful properties.

Theorem A.8.1 *Properties of eigenvalues. Let λ_i and $\boldsymbol{h}_i$, $i = 1, ..., k$, denote the k eigenvalues and eigenvectors of a square matrix $\boldsymbol{A}$.*

1. $\det(\boldsymbol{A}) = \prod_{i=1}^{k} \lambda_i$
2. $\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} \lambda_i$
3. $\boldsymbol{A}$ is non-singular if and only if all its eigenvalues are non-zero.
4. If $\boldsymbol{A}$ has distinct eigenvalues, there exists a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{A} = \boldsymbol{P}^{-1}\boldsymbol{\Lambda}\boldsymbol{P}$ and $\boldsymbol{P}\boldsymbol{A}\boldsymbol{P}^{-1} = \boldsymbol{\Lambda}$.
5. The non-zero eigenvalues of $\boldsymbol{A}\boldsymbol{B}$ and $\boldsymbol{B}\boldsymbol{A}$ are identical.
6. If $\boldsymbol{B}$ is non-singular then $\boldsymbol{A}$ and $\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{B}$ have the same eigenvalues.
7. If $\boldsymbol{A}\boldsymbol{h} = \boldsymbol{h}\lambda$ then $(\boldsymbol{I} - \boldsymbol{A}) = \boldsymbol{h}(1 - \lambda)$. So $\boldsymbol{I} - \boldsymbol{A}$ has the eigenvalue $1 - \lambda$ and associated eigenvector $\boldsymbol{h}$.

Most eigenvalue applications in econometrics concern the case where the matrix $\boldsymbol{A}$ is real and symmetric. In this case all eigenvalues of $\boldsymbol{A}$ are real and its eigenvectors are mutually orthogonal. Thus $\boldsymbol{H}$ is orthonormal so $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_k$ and $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{I}_k$. When the eigenvalues are all real it is conventional to write them in decending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$.

The following is a very important property of real symmetric matrices, which follows directly from the equations $\boldsymbol{A}\boldsymbol{H} = \boldsymbol{H}\boldsymbol{\Lambda}$ and $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_k$.

Spectral Decomposition. If $\boldsymbol{A}$ is a $k \times k$ real symmetric matrix, then $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}$ contains the eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix with the eigenvalues on the diagaonal. The eigenvalues are all real and the eigenvector matrix satisfies $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_k$. The decomposition can be alternatively written as $\boldsymbol{H}'\boldsymbol{A}\boldsymbol{H} = \boldsymbol{\Lambda}$.

If $\boldsymbol{A}$ is real, symmetric, and invertible, then by the spectral decomposition and the properties of orthonormal matrices, $\boldsymbol{A}^{-1} = \boldsymbol{H}'^{-1}\boldsymbol{\Lambda}^{-1}\boldsymbol{H}^{-1} = \boldsymbol{H}\boldsymbol{\Lambda}^{-1}\boldsymbol{H}'$. Thus the columns of $\boldsymbol{H}$ are also the eigenvectors of $\boldsymbol{A}^{-1}$, and its eigenvalues are λ_1^{-1}, λ_2^{-1}, ..., λ_k^{-1}.

A.9 Positive Definite Matrices

We say that a $k \times k$ real symmetric square matrix $\boldsymbol{A}$ is **positive semi-definite** if for all $\boldsymbol{c} \neq \boldsymbol{0}$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \geq 0$. This is written as $\boldsymbol{A} \geq 0$. We say that $\boldsymbol{A}$ is **positive definite** if for all $\boldsymbol{c} \neq \boldsymbol{0}$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} > 0$. This is written as $\boldsymbol{A} > 0$.

Some properties include:

Theorem A.9.1 *Properties of positive semi-definite matrices*

1. If $\boldsymbol{A} = \boldsymbol{G}'\boldsymbol{B}\boldsymbol{G}$ with $\boldsymbol{B} \geq 0$ and some matrix $\boldsymbol{G}$**,** then $\boldsymbol{A}$ is positive semi-definite. (For any $\boldsymbol{c} \neq \boldsymbol{0}$, $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} = \boldsymbol{\alpha}'\boldsymbol{B}\boldsymbol{\alpha} \geq 0$ where $\boldsymbol{\alpha} = \boldsymbol{G}\boldsymbol{c}$.) If $\boldsymbol{G}$ has full column rank and $\boldsymbol{B} > 0$, then $\boldsymbol{A}$ is positive definite.
2. If $\boldsymbol{A}$ is positive definite, then $\boldsymbol{A}$ is non-singular and $\boldsymbol{A}^{-1}$ exists. Furthermore, $\boldsymbol{A}^{-1} > 0$.
3. $\boldsymbol{A} > 0$ if and only if it is symmetric and all its eigenvalues are positive.
4. By the spectral decomposition, $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_k$ and $\boldsymbol{\Lambda}$ is diagonal with non-negative diagonal elements. All diagonal elements of $\boldsymbol{\Lambda}$ are strictly positive if (and only if) $\boldsymbol{A} > 0$.
5. The rank of $\boldsymbol{A}$ equals the number of strictly positive eigenvalues.
6. If $\boldsymbol{A} > 0$ then $\boldsymbol{A}^{-1} = \boldsymbol{H}\boldsymbol{\Lambda}^{-1}\boldsymbol{H}'$.
7. If $\boldsymbol{A} \geq 0$ and $\operatorname{rank}(\boldsymbol{A}) = r \leq k$ then the Moore-Penrose generalized inverse of $\boldsymbol{A}$ is $\boldsymbol{A}^{-} = \boldsymbol{H}\boldsymbol{\Lambda}^{-}\boldsymbol{H}'$ where $\boldsymbol{\Lambda}^{-} = \operatorname{diag}\left(\lambda_1^{-1}, \lambda_2^{-1}, ..., \lambda_r^{-1}, 0, ..., 0\right)$.
8. If $\boldsymbol{A} \geq 0$ we can find a matrix $\boldsymbol{B}$ such that $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}'$. We call $\boldsymbol{B}$ a **matrix square root** of $\boldsymbol{A}$ and is typically written as $\boldsymbol{B} = \boldsymbol{A}^{1/2}$. The matrix $\boldsymbol{B}$ need not be unique. One matrix square root is obtained using the spectral decomposition $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$. Then $\boldsymbol{B} = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}\boldsymbol{H}'$ is itself symmetric and positive definite and satisfies $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{B}$. Another matrix square root is the Cholesky decomposition, described in Section A.14.

A.10 Generalized Eigenvalues

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $k \times k$ matrices. The generalized characteristic equation is

$$\det(\mu\boldsymbol{B} - \boldsymbol{A}) = 0.$$

The solutions μ are known as **generalized eigenvalues** of $\boldsymbol{A}$ with respect to $\boldsymbol{B}$. Associated with each generalized eigenvalue μ is a **generalized eigenvector** $\boldsymbol{v}$ which satisfies

$$\boldsymbol{A}\boldsymbol{v} = \boldsymbol{B}\boldsymbol{v}\mu.$$

They are typically normalized so that $\boldsymbol{v}'\boldsymbol{B}\boldsymbol{v} = 1$ and thus $\mu = \boldsymbol{v}'\boldsymbol{A}\boldsymbol{v}$.

A matrix expression is

$$\boldsymbol{A}\boldsymbol{V} = \boldsymbol{B}\boldsymbol{V}\boldsymbol{M}$$

where $\boldsymbol{M} = \operatorname{diag}\{\mu_1, ..., \mu_k\}$.

If $\boldsymbol{A}$ and $\boldsymbol{B}$ are real and symmetric then the generalized eigenvalues are real.

Suppose in addition that $\boldsymbol{B}$ is invertible. Then the generalized eigenvalues of $\boldsymbol{A}$ with respect to $\boldsymbol{B}$ are equal to the eigenvalues of $\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2\prime}$. The generalized eigenvectors $\boldsymbol{V}$ of $\boldsymbol{A}$ with respect to $\boldsymbol{B}$ are related to the eigenvectors $\boldsymbol{H}$ of $\boldsymbol{B}^{-1/2}\boldsymbol{A}\boldsymbol{B}^{-1/2\prime}$ by the relationship $\boldsymbol{V} = \boldsymbol{B}^{-1/2\prime}\boldsymbol{H}$. This implies $\boldsymbol{V}'\boldsymbol{B}\boldsymbol{V} = \boldsymbol{I}_k$. Thus the generalized eigenvectors are orthogonalized with respect to the matrix $\boldsymbol{B}$.

If $\boldsymbol{A}\boldsymbol{v} = \boldsymbol{B}\boldsymbol{v}\mu$ then $(\boldsymbol{B} - \boldsymbol{A})\,\boldsymbol{v} = \boldsymbol{B}\boldsymbol{v}(1-\mu)$. So a generalized eigenvalue of $\boldsymbol{B} - \boldsymbol{A}$ with respect to $\boldsymbol{B}$ is $1-\mu$ with associated eigenvector $\boldsymbol{v}$.

Generalized eigenvalue equations have an interesting dual property. The following is based on Lemma A.9 of Johansen (1995).

Theorem A.10.1 *Suppose that $\boldsymbol{B}$ and $\boldsymbol{C}$ are invertible $p \times p$ and $r \times r$ matrices, respectively, and $\boldsymbol{A}$ is $p \times r$. Then the generalized eigenvalue problems*

$$\det\left(\mu\boldsymbol{B} - \boldsymbol{A}\boldsymbol{C}^{-1}\boldsymbol{A}'\right) = 0 \tag{A.5}$$

and

$$\det\left(\mu\boldsymbol{C} - \boldsymbol{A}'\boldsymbol{B}^{-1}\boldsymbol{A}\right) = 0 \tag{A.6}$$

have the same non-zero generalized eigenvalues. Furthermore, for any such generalized eigenvalue μ, if $\boldsymbol{v}$ and $\boldsymbol{w}$ are the associated generalized eigenvectors of (A.5) and (A.6), then

$$\boldsymbol{w} = \mu^{-1/2}\boldsymbol{C}^{-1}\boldsymbol{A}'\boldsymbol{v}. \tag{A.7}$$

Proof:. Let $\mu \neq 0$ be an eigenvalue of (A.5). Then using Theorem A.7.1.8

$$\begin{aligned} 0 &= \det\left(\mu\boldsymbol{B} - \boldsymbol{A}\boldsymbol{C}^{-1}\boldsymbol{A}'\right) \\ &= \frac{\det(\mu\boldsymbol{B})}{\det(\boldsymbol{C})}\det\left(\boldsymbol{C} - \boldsymbol{A}'(\mu\boldsymbol{B})^{-1}\boldsymbol{A}\right) \\ &= \frac{\det(\boldsymbol{B})}{\det(\boldsymbol{C})}\det\left(\mu\boldsymbol{C} - \boldsymbol{A}'\boldsymbol{B}^{-1}\boldsymbol{A}\right). \end{aligned}$$

Since $\det(\boldsymbol{B})/\det(\boldsymbol{C}) \neq 0$ this implies (A.7) holds. Hence μ is an eigenvalue of (A.6), as claimed.

We next show that (A.7) is an eigenvector of (A.6). Note that the solutions to (A.5) and (A.6) satisfy

$$\boldsymbol{B}\boldsymbol{v}\mu = \boldsymbol{A}\boldsymbol{C}^{-1}\boldsymbol{A}'\boldsymbol{v} \tag{A.8}$$

and

$$\boldsymbol{C}\boldsymbol{w}\mu = \boldsymbol{A}'\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{w} \tag{A.9}$$

and are normalized so that $\boldsymbol{v}'\boldsymbol{B}\boldsymbol{v} = 1$ and $\boldsymbol{w}'\boldsymbol{C}\boldsymbol{w} = 1$. We show that (A.7) satisfies (A.9). Using (A.7), we find that the left-side of (A.9) equals

$$\boldsymbol{C}\left(\mu^{-1/2}\boldsymbol{C}^{-1}\boldsymbol{A}'\right)\mu = \boldsymbol{A}'\mu^{1/2} = \boldsymbol{A}'\boldsymbol{B}^{-1}\boldsymbol{B}\boldsymbol{v}\mu^{1/2} = \boldsymbol{A}'\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{C}^{-1}\boldsymbol{A}'\boldsymbol{v}\mu^{-1/2} = \boldsymbol{A}'\boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{w}$$

The third equality is (A.8) and the final is (A.7). This shows that (A.9) holds and thus (A.7) is an eigenvector of (A.6) as stated. ■

A.11 Extrema of Quadratic Forms

The extrema of quadratic forms in real symmetric matrices can be conveniently be written in terms of eigenvalues and eigenvectors.

Let $\boldsymbol{A}$ denote a $k \times k$ real symmetric matrix. Let $\lambda_1 \geq \cdots \geq \lambda_k$ be the ordered eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{h}_1, ..., \boldsymbol{h}_k$ the associated ordered eigenvectors.

We start with results for the extrema of $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$. Throughout this Section, when we refer to the "solution" of an extremum problem, it is the solution to the normalized expression.

- $\max_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \max_{\boldsymbol{x}} \dfrac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \lambda_1$. The solution is $\boldsymbol{x} = \boldsymbol{h}_1$. (That is, the maximizer of $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ over $\boldsymbol{x}'\boldsymbol{x} = 1$.)

- $\min_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \min_{\boldsymbol{x}} \dfrac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \lambda_k$. The solution is $\boldsymbol{x} = \boldsymbol{h}_k$.

Multivariate generalizations can involve either the trace or the determinant.

- $\max_{\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{I}_\ell} \operatorname{tr}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \max_{\boldsymbol{X}} \operatorname{tr}\left((\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})\right) = \sum_{i=1}^{\ell} \lambda_i$.
  The solution is $\boldsymbol{X} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_\ell]$.
- $\min_{\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{I}_\ell} \operatorname{tr}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \min_{\boldsymbol{X}} \left((\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})\right) = \sum_{i=1}^{\ell} \lambda_{k-i+1}$.
  The solution is $\boldsymbol{X} = [\boldsymbol{h}_{k-\ell+1}, ..., \boldsymbol{h}_k]$.

For a proof, see Theorem 11.13 of Magnus and Neudecker (1988).

Suppose as well that $\boldsymbol{A} > 0$ with ordered eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ and eigenvectors $[\boldsymbol{h}_1, ..., \boldsymbol{h}_k]$

- $\max_{\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{I}_\ell} \det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \max_{\boldsymbol{X}} \dfrac{\det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{X})} = \prod_{i=1}^{\ell} \lambda_i$. The solution is $\boldsymbol{X} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_\ell]$.
- $\min_{\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{I}_\ell} \det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \min_{\boldsymbol{X}} \dfrac{\det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{X})} = \prod_{i=1}^{\ell} \lambda_{k-i+1}$. The solution is $\boldsymbol{X} = [\boldsymbol{h}_{k-\ell+1}, ..., \boldsymbol{h}_k]$.
- $\max_{\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{I}_\ell} \det(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X}) = \max_{\boldsymbol{X}} \dfrac{\det(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{X})} = \prod_{i=1}^{\ell} (1-\lambda_{k-i+1})$. The solution is $\boldsymbol{X} = [\boldsymbol{h}_{k-\ell+1}, ..., \boldsymbol{h}_k]$.
- $\min_{\boldsymbol{X}'\boldsymbol{X}=\boldsymbol{I}_\ell} \det(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X}) = \min_{\boldsymbol{X}} \dfrac{\det(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{X})} = \prod_{i=1}^{\ell} (1-\lambda_i)$. The solution is $\boldsymbol{X} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_\ell]$.

For a proof, see Theorem 11.15 of Magnus and Neudecker (1988).

We can extend the above results to incorporate generalized eigenvalue equations.

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $k \times k$ real symmetric matrices with $\boldsymbol{B} > 0$. Let $\mu_1 \geq \cdots \geq \mu_k$ be the ordered generalized eigenvalues of $\boldsymbol{A}$ with respect to $\boldsymbol{B}$ and $\boldsymbol{v}_1, ..., \boldsymbol{v}_k$ the associated ordered eigenvectors.

- $\max_{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \max_{\boldsymbol{x}} \dfrac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}} = \mu_1$. The solution is $\boldsymbol{x} = \boldsymbol{v}_1$.
- $\min_{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = \min_{\boldsymbol{x}} \dfrac{\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{B}\boldsymbol{x}} = \mu_k$. The solution is $\boldsymbol{x} = \boldsymbol{v}_k$.
- $\max_{\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}=\boldsymbol{I}_\ell} \operatorname{tr}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \max_{\boldsymbol{X}} \operatorname{tr}\left((\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})\right) = \sum_{i=1}^{\ell} \mu_i$.
  The solution is $\boldsymbol{X} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_\ell]$.
- $\min_{\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}=\boldsymbol{I}_\ell} \operatorname{tr}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \min_{\boldsymbol{X}} \operatorname{tr}\left((\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})\right) = \sum_{i=1}^{\ell} \mu_{k-i+1}$.
  The solution is $\boldsymbol{X} = [\boldsymbol{v}_{k-\ell+1}, ..., \boldsymbol{v}_k]$.

Suppose as well that $\boldsymbol{A} > 0$.

- $\max_{\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}=\boldsymbol{I}_\ell} \det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \max_{\boldsymbol{X}} \frac{\det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X})} = \prod_{i=1}^{\ell} \mu_i.$

  The solution is $\boldsymbol{X} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_\ell]$.

- $\min_{\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}=\boldsymbol{I}_\ell} \det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X}) = \min_{\boldsymbol{X}} \frac{\det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X})} = \prod_{i=1}^{\ell} \mu_{k-i+1}.$

  The solution is $\boldsymbol{X} = [\boldsymbol{v}_{k-\ell+1}, ..., \boldsymbol{v}_k]$.

- $\max_{\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}=\boldsymbol{I}_\ell} \det(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X}) = \max_{\boldsymbol{X}} \frac{\det(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X})} = \prod_{i=1}^{\ell} (1-\mu_{k-i+1}).$

  The solution is $\boldsymbol{X} = [\boldsymbol{v}_{k-\ell+1}, ..., \boldsymbol{v}_k]$.

- $\min_{\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X}=\boldsymbol{I}_\ell} \det(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X}) = \min_{\boldsymbol{X}} \frac{\det(\boldsymbol{X}'(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X})} = \prod_{i=1}^{\ell} (1-\mu_i).$

  The solution is $\boldsymbol{X} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_\ell]$..

By change-of-variables, we can re-express one eigenvalue problem in terms of another. For example, let $\boldsymbol{A} > 0$, $\boldsymbol{B} > 0$, and $\boldsymbol{C} > 0$. Then

$$\max_{\boldsymbol{X}} \frac{\det(\boldsymbol{X}'\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}\boldsymbol{X})} = \max_{\boldsymbol{X}} \frac{\det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X})}$$

and

$$\min_{\boldsymbol{X}} \frac{\det(\boldsymbol{X}'\boldsymbol{C}\boldsymbol{A}\boldsymbol{C}\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}\boldsymbol{X})} = \min_{\boldsymbol{X}} \frac{\det(\boldsymbol{X}'\boldsymbol{A}\boldsymbol{X})}{\det(\boldsymbol{X}'\boldsymbol{B}\boldsymbol{X})}.$$

A.12 Idempotent Matrices

A $k \times k$ square matrix $\boldsymbol{A}$ is **idempotent** if $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. When $k = 1$ the only idempotent numbers are 1 and 0. For $k > 1$ there are many possibilities. For example, the following matrix is idempotent

$$\boldsymbol{A} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix}.$$

If $\boldsymbol{A}$ is idempotent and symmetric with rank r, then it has r eigenvalues which equal 1 and $k-r$ eigenvalues which equal 0. To see this, by the spectral decomposition we can write $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}$ is orthonormal and $\boldsymbol{\Lambda}$ contains the eigenvalues. Then

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}' = \boldsymbol{H}\boldsymbol{\Lambda}^2\boldsymbol{H}'.$$

We deduce that $\boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda}$ and $\lambda_i^2 = \lambda_i$ for $i = 1, ..., k$. Hence each λ_i must equal either 0 or 1. Since the rank of $\boldsymbol{A}$ is r, and the rank equals the number of positive eigenvalues, it follows that

$$\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}_{k-r} \end{bmatrix}.$$

Thus the spectral decomposition of an idempotent matrix $\boldsymbol{A}$ takes the form

$$\boldsymbol{A} = \boldsymbol{H} \begin{bmatrix} \boldsymbol{I}_r & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}_{k-r} \end{bmatrix} \boldsymbol{H}' \tag{A.10}$$

with $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_k$. Additionally, $\mathrm{tr}(\boldsymbol{A}) = \mathrm{rank}(\boldsymbol{A})$ and $\boldsymbol{A}$ is positive semi-definite.

If $\boldsymbol{A}$ is idempotent and symmetric with rank $r < k$ then it does not possess an inverse, but its Moore-Penrose generalized inverse takes the simple form $\boldsymbol{A}^- = \boldsymbol{A}$. This can be verified by checking the conditions for the Moore-Penrose generalized inverse , for example $\boldsymbol{A}\boldsymbol{A}^-\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$.

If $\boldsymbol{A}$ is idempotent then $\boldsymbol{I} - \boldsymbol{A}$ is also idempotent.

One useful fact is that if $\boldsymbol{A}$ is idempotent then for any conformable vector $\boldsymbol{c}$,

$$\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \leq \boldsymbol{c}'\boldsymbol{c} \tag{A.11}$$

$$\boldsymbol{c}'(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{c} \leq \boldsymbol{c}'\boldsymbol{c} \tag{A.12}$$

To see this, note that

$$\boldsymbol{c}'\boldsymbol{c} = \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} + \boldsymbol{c}'(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{c}.$$

Since $\boldsymbol{A}$ and $\boldsymbol{I} - \boldsymbol{A}$ are idempotent, they are both positive semi-definite, so both $\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c}$ and $\boldsymbol{c}'(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{c}$ are non-negative. Thus they must satisfy (A.11)-(A.12).

A.13 Singular Values

The singular values of a $k \times r$ real matrix $\boldsymbol{A}$ are the positive square roots of the eigenvalues of $\boldsymbol{A}'\boldsymbol{A}$. Thus for $j = 1, ..., r$

$$s_j = \sqrt{\lambda_j(\boldsymbol{A}'\boldsymbol{A})}$$

Since $\boldsymbol{A}'\boldsymbol{A}$ is positive semi-definite, its eigenvalues are non-negative. Thus singular values are always real and non-negative.

The non-zero singular values of $\boldsymbol{A}$ and $\boldsymbol{A}'$ are the same.

When $\boldsymbol{A}$ is positive semi-definite then the singular values of $\boldsymbol{A}$ correspond to its eigenvalues.

The singular value decomposition of a $k \times r$ real matrix $\boldsymbol{A}$ takes the form $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}'$ where $\boldsymbol{U}$ is $k \times k$, $\boldsymbol{\Lambda}$ is $k \times r$ and $\boldsymbol{V}$ is $r \times r$, with $\boldsymbol{U}$ and $\boldsymbol{V}$ orthonormal ($\boldsymbol{U}'\boldsymbol{U} = \boldsymbol{I}_k$ and $\boldsymbol{V}'\boldsymbol{V} = \boldsymbol{I}_r$) and $\boldsymbol{\Lambda}$ is a diagonal matrix with the singular values of $\boldsymbol{A}$ on the diagonal.

It is convention to write the singular values in decending order $s_1 \geq s_2 \geq \cdots \geq s_r$.

A.14 Cholesky Decomposition

For a $k \times k$ positive definite matrix $\boldsymbol{A}$, its **Cholesky decomposition** takes the form

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}'$$

where $\boldsymbol{L}$ is **lower triangular**, and thus takes the form

$$\boldsymbol{L} = \begin{bmatrix} L_{11} & 0 & \cdots & 0 \\ L_{21} & L_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ L_{k1} & L_{k2} & \cdots & L_{kk} \end{bmatrix}.$$

The diagonal elements of $\boldsymbol{L}$ are all strictly positive.

The Cholesky decomposition is unique (for positive definite $\boldsymbol{A}$). One intuition is that the matrices $\boldsymbol{A}$ and $\boldsymbol{L}$ each have $k(k+1)/2$ free elements.

The decomposition is very useful for a range of computations, especially when a matrix square root is required. Algorithms for computation are available in standard packages (for example, `chol` in either MATLAB or R).

Lower triangular matrices such as $\boldsymbol{L}$ have special properties. One is that its determinant equals the product of the diagonal elements.

Proofs of uniqueness are algorithmic. Here is one such argument for the case $k = 3$. Write out

$$\begin{bmatrix} A_{11} & A_{21} & A_{31} \\ A_{21} & A_{22} & A_{32} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} = \boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}' = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix} \begin{bmatrix} L_{11} & L_{21} & L_{31} \\ 0 & L_{22} & L_{32} \\ 0 & 0 & L_{33} \end{bmatrix}$$
$$= \begin{bmatrix} L_{11}^2 & L_{11}L_{21} & L_{11}L_{31} \\ L_{11}L_{21} & L_{21}^2 + L_{22}^2 & L_{31}L_{21} + L_{32}L_{22} \\ L_{11}L_{31} & L_{31}L_{21} + L_{32}L_{22} & L_{31}^2 + L_{32}^2 + L_{33}^2 \end{bmatrix}$$

There are six equations, six knowns (the elements of $\boldsymbol{A}$) and six unknowns (the elements of $\boldsymbol{L}$). We can solve for the latter by starting with the first column, moving from top to bottom. The first element has the simple solution

$$L_{11} = \sqrt{A_{11}}.$$

This has a real solution since $A_{11} > 0$. Moving down, since L_{11} is known, for the entries beneath L_{11} we solve and find

$$L_{21} = \frac{A_{21}}{L_{11}} = \frac{A_{21}}{\sqrt{A_{11}}}$$
$$L_{31} = \frac{A_{31}}{L_{11}} = \frac{A_{31}}{\sqrt{A_{11}}}$$

Next we move to the second column. We observe that L_{21} is known. Then we solve for L_{22}

$$L_{22} = \sqrt{A_{22} - L_{21}^2} = \sqrt{A_{22} - \frac{A_{21}^2}{A_{11}}}.$$

This has a real solution since $\boldsymbol{A} > 0$. Then since L_{22} is known we can move down the column to find

$$L_{32} = \frac{A_{32} - L_{31}L_{21}}{L_{22}} = \frac{A_{32} - \frac{A_{31}A_{21}}{A_{11}}}{\sqrt{A_{22} - \frac{A_{21}^2}{A_{11}}}}.$$

Finally we take the third column. All elements except L_{33} are known. So we solve to find

$$L_{33} = \sqrt{A_{33} - L_{31}^2 - L_{32}^2} = \sqrt{A_{33} - \frac{A_{31}^2}{A_{11}} - \frac{\left(A_{32} - \frac{A_{31}A_{21}}{A_{11}}\right)^2}{A_{22} - \frac{A_{21}^2}{A_{11}}}}.$$

A.15 Matrix Calculus

Let $\boldsymbol{x} = (x_1, ..., x_k)'$ be $k \times 1$ and $g(\boldsymbol{x}) = g(x_1, ..., x_k) : \mathbb{R}^k \to \mathbb{R}$. The vector derivative is

$$\frac{\partial}{\partial \boldsymbol{x}} g(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(\boldsymbol{x}) \\ \vdots \\ \frac{\partial}{\partial x_k} g(\boldsymbol{x}) \end{pmatrix}$$

and

$$\frac{\partial}{\partial \boldsymbol{x}'} g(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial}{\partial x_1} g(\boldsymbol{x}) & \cdots & \frac{\partial}{\partial x_k} g(\boldsymbol{x}) \end{pmatrix}.$$

Some properties are now summarized.

Theorem A.15.1 *Properties of matrix derivatives*

1. $\frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{a}'\boldsymbol{x}) = \frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{x}'\boldsymbol{a}) = \boldsymbol{a}$

2. $\frac{\partial}{\partial \boldsymbol{x}'}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}$
3. $\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}')\boldsymbol{x}$
4. $\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A} + \boldsymbol{A}'$
5. $\frac{\partial}{\partial \boldsymbol{A}} \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}) = \boldsymbol{B}'$
6. $\frac{\partial}{\partial \boldsymbol{A}} \log\det(\boldsymbol{A}) = \left(\boldsymbol{A}^{-}\right)'$

The final two results require some justification. Recall from Section A.5 that we can write out explicitly

$$\operatorname{tr}(\boldsymbol{B}\boldsymbol{A}) = \sum_i \sum_j a_{ij} b_{ji}.$$

Thus if we take the derivative with respect to a_{ij} we find

$$\frac{\partial}{\partial a_{ij}} \operatorname{tr}(\boldsymbol{B}\boldsymbol{A}) = b_{ji}.$$

which is the ij^{th} element of $\boldsymbol{B}'$, establishing part 5.

For part 6, recall Laplace's expansion

$$\det \boldsymbol{A} = \sum_{j=1}^{k} a_{ij} C_{ij}.$$

where C_{ij} is the ij^{th} cofactor of $\boldsymbol{A}$. Set $\boldsymbol{C} = (C_{ij})$. Observe that C_{ij} for $j = 1, ..., k$ are not functions of a_{ij}. Thus the derivative with respect to a_{ij} is

$$\frac{\partial}{\partial a_{ij}} \log\det(\boldsymbol{A}) = (\det \boldsymbol{A})^{-1} \frac{\partial}{\partial a_{ij}} \det \boldsymbol{A} = (\det \boldsymbol{A})^{-1} C_{ij}$$

Together this implies

$$\frac{\partial}{\partial \boldsymbol{A}} \log\det(\boldsymbol{A}) = (\det \boldsymbol{A})^{-1} \boldsymbol{C} = \boldsymbol{A}^{-1}$$

where the second equality is Theorem A.7.1.12.

A.16 Kronecker Products and the Vec Operator

Let $\boldsymbol{A} = [\boldsymbol{a}_1 \ \boldsymbol{a}_2 \ \cdots \ \boldsymbol{a}_n]$ be $m \times n$. The **vec** of $\boldsymbol{A}$, denoted by $\operatorname{vec}(\boldsymbol{A})$, is the $mn \times 1$ vector

$$\operatorname{vec}(\boldsymbol{A}) = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_n \end{pmatrix}.$$

Let $\boldsymbol{A} = (a_{ij})$ be an $m \times n$ matrix and let $\boldsymbol{B}$ be any matrix. The **Kronecker product** of $\boldsymbol{A}$ and $\boldsymbol{B}$, denoted $\boldsymbol{A} \otimes \boldsymbol{B}$, is the matrix

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

Some important properties are now summarized. These results hold for matrices for which all matrix multiplications are conformable.

Theorem A.16.1 *Properties of the Kronecker product*

1. $(\boldsymbol{A}+\boldsymbol{B})\otimes\boldsymbol{C}=\boldsymbol{A}\otimes\boldsymbol{C}+\boldsymbol{B}\otimes\boldsymbol{C}$
2. $(\boldsymbol{A}\otimes\boldsymbol{B})(\boldsymbol{C}\otimes\boldsymbol{D})=\boldsymbol{AC}\otimes\boldsymbol{BD}$
3. $\boldsymbol{A}\otimes(\boldsymbol{B}\otimes\boldsymbol{C})=(\boldsymbol{A}\otimes\boldsymbol{B})\otimes C$
4. $(\boldsymbol{A}\otimes\boldsymbol{B})'=\boldsymbol{A}'\otimes\boldsymbol{B}'$
5. $\operatorname{tr}(\boldsymbol{A}\otimes\boldsymbol{B})=\operatorname{tr}(\boldsymbol{A})\operatorname{tr}(\boldsymbol{B})$
6. If $\boldsymbol{A}$ is $m\times m$ and $\boldsymbol{B}$ is $n\times n$, $\det(\boldsymbol{A}\otimes\boldsymbol{B})=(\det(\boldsymbol{A}))^{n}(\det(\boldsymbol{B}))^{m}$
7. $(\boldsymbol{A}\otimes\boldsymbol{B})^{-1}=\boldsymbol{A}^{-1}\otimes\boldsymbol{B}^{-1}$
8. If $\boldsymbol{A}>0$ and $\boldsymbol{B}>0$ then $\boldsymbol{A}\otimes\boldsymbol{B}>0$
9. $\operatorname{vec}(\boldsymbol{ABC})=(\boldsymbol{C}'\otimes\boldsymbol{A})\operatorname{vec}(\boldsymbol{B})$
10. $\operatorname{tr}(\boldsymbol{ABCD})=\operatorname{vec}(\boldsymbol{D}')'(\boldsymbol{C}'\otimes\boldsymbol{A})\operatorname{vec}(\boldsymbol{B})$

A.17 Vector Norms

Given any vector space V (such as Euclidean space $\mathbb{R}^m$) a **norm** on V is a function $\rho:V\to\mathbb{R}$ with the properties

1. $\rho(c\boldsymbol{a})=|c|\,\rho(\boldsymbol{a})$ for any complex number c and $\boldsymbol{a}\in V$
2. $\rho(\boldsymbol{a}+\boldsymbol{b})\leq\rho(\boldsymbol{a})+\rho(\boldsymbol{b})$
3. If $\rho(\boldsymbol{a})=0$ then $\boldsymbol{a}=\boldsymbol{0}$

A seminorm on V is a function which satisfies the first two properties. The second property is known as the triangle inequality, and it is the one property which typically needs a careful demonstration (as the other two properties typically hold by inspection).

The typical norm used for Euclidean space $\mathbb{R}^m$ is the **Euclidean norm**

$$\begin{aligned}\|\boldsymbol{a}\| &= \left(\boldsymbol{a}'\boldsymbol{a}\right)^{1/2}\\ &= \left(\sum_{i=1}^{m}a_i^2\right)^{1/2}.\end{aligned}$$

An alternative norm is the $p-$norm (for $p\geq1$)

$$\|\boldsymbol{a}\|_p=\left(\sum_{i=1}^{m}|a_i|^p\right)^{1/p}.$$

Special cases include the Euclidean norm ($p=2$), the $1-$norm

$$\|\boldsymbol{a}\|_1=\sum_{i=1}^{m}|a_i|$$

and the sup-norm

$$\|\boldsymbol{a}\|_\infty=\max(|a_1|,...,|a_m|).$$

For real numbers ($m=1$) these norms coincide.

Some standard inequalities for Euclidean space are now given. The Minkowski inequality given below establishes that any p-norm with $p \geq 1$ (including the Euclidean norm) satisfies the triangle inequality and is thus a valid norm.

Jensen's Inequality. If $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is convex, then for any non-negative weights a_j such that $\sum_{j=1}^{m} a_j = 1$, and any real numbers x_j

$$g\left(\sum_{j=1}^{m} a_j x_j\right) \leq \sum_{j=1}^{m} a_j g\left(x_j\right). \tag{A.13}$$

In particular, setting $a_j = 1/m$, then

$$g\left(\frac{1}{m}\sum_{j=1}^{m} x_j\right) \leq \frac{1}{m}\sum_{j=1}^{m} g\left(x_j\right). \tag{A.14}$$

If $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is concave then the inequalities in (A.13) and (A.14) are reversed.

Weighted Geometric Mean Inequality. For any non-negative real weights a_j such that $\sum_{j=1}^{m} a_j = 1$, and any non-negative real numbers x_j

$$x_1^{a_1} x_2^{a_2} \cdots x_m^{a_m} \leq \sum_{j=1}^{m} a_j x_j \tag{A.15}$$

Loève's c_r Inequality. For $r > 0$,

$$\left|\sum_{j=1}^{m} a_j\right|^r \leq c_r \sum_{j=1}^{m} |a_j|^r \tag{A.16}$$

where $c_r = 1$ when $r \leq 1$ and $c_r = m^{r-1}$ when $r \geq 1$.

c_2 Inequality. For any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$(\boldsymbol{a}+\boldsymbol{b})'(\boldsymbol{a}+\boldsymbol{b}) \leq 2\boldsymbol{a}'\boldsymbol{a} + 2\boldsymbol{b}'\boldsymbol{b} \tag{A.17}$$

Hölder's Inequality. If $p > 1$, $q > 1$, and $1/p + 1/q = 1$, then for any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\sum_{j=1}^{m} |a_j b_j| \leq \|\boldsymbol{a}\|_p \|\boldsymbol{b}\|_q \tag{A.18}$$

Minkowski's Inequality. For any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, if $p \geq 1$, then

$$\|\boldsymbol{a}+\boldsymbol{b}\|_p \leq \|\boldsymbol{a}\|_p + \|\boldsymbol{b}\|_p \tag{A.19}$$

Schwarz Inequality. For any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\left|\boldsymbol{a}'\boldsymbol{b}\right| \leq \|\boldsymbol{a}\|\,\|\boldsymbol{b}\|. \tag{A.20}$$

Proof of Jensen's Inequality (A.13). By the definition of convexity, for any $\lambda \in [0,1]$

$$g\left(\lambda x_1 + (1-\lambda)x_2\right) \leq \lambda g\left(x_1\right) + (1-\lambda)\, g\left(x_2\right). \tag{A.21}$$

This implies

$$g\left(\sum_{j=1}^{m} a_j x_j\right) = g\left(a_1 x_1 + (1-a_1)\sum_{j=2}^{m} \frac{a_j}{1-a_1} x_j\right)$$
$$\leq a_1 g(x_1) + (1-a_1)\, g\left(\sum_{j=2}^{m} b_j x_j\right)$$

where $b_j = a_j/(1-a_1)$ and $\sum_{j=2}^{m} b_j = 1$. By another application of (A.21) this is bounded by

$$a_1 g(x_1) + (1-a_1)\left(b_2 g(x_2) + (1-b_2) g\left(\sum_{j=2}^{m} c_j x_j\right)\right)$$
$$= a_1 g(x_1) + a_2 g(x_2) + (1-a_1)(1-b_2) g\left(\sum_{j=2}^{m} c_j x_j\right)$$

where $c_j = b_j/(1-b_2)$. By repeated application of (A.21) we obtain (A.13). ■

Proof of Weighted Geometric Mean Inequality. Since the logarithm is strictly concave, by Jensen's inequality

$$\log\left(x_1^{a_1} x_2^{a_2} \cdots x_m^{a_m}\right) = \sum_{j=1}^{m} a_j \log x_j \leq \log\left(\sum_{j=1}^{m} a_j x_j\right).$$

Applying the exponential yields (A.15). ■

Proof of Loève's c_r Inequality. For $r \geq 1$ this is simply a rewriting of the finite form Jensen's inequality (A.14) with $g(u) = u^r$. For $r < 1$, define $b_j = |a_j| \,/\, \left(\sum_{j=1}^{m} |a_j|\right)$. The facts that $0 \leq b_j \leq 1$ and $r < 1$ imply $b_j \leq b_j^r$ and thus

$$1 = \sum_{j=1}^{m} b_j \leq \sum_{j=1}^{m} b_j^r$$

which implies

$$\left(\sum_{j=1}^{m} |a_j|\right)^r \leq \sum_{j=1}^{m} |a_j|^r.$$

The proof is completed by observing that

$$\left(\sum_{j=1}^{m} a_j\right)^r \leq \left(\sum_{j=1}^{m} |a_j|\right)^r.$$

■

Proof of c_2 Inequality. By the c_r inequality, $(a_j + b_j)^2 \leq 2a_j^2 + 2b_j^2$. Thus

$$(\boldsymbol{a}+\boldsymbol{b})'(\boldsymbol{a}+\boldsymbol{b}) = \sum_{j=1}^{m} (a_j + b_j)^2$$
$$\leq 2\sum_{j=1}^{m} a_j^2 + 2\sum_{j=1}^{m} b_j^2$$
$$= 2\boldsymbol{a}'\boldsymbol{a} + 2\boldsymbol{b}'\boldsymbol{b}$$

■

Proof of Hölder's Inequality. Set $u_j = |a_j|^p / \|\boldsymbol{a}\|_p^p$ and $u_j = |b_j|^q / \|\boldsymbol{b}\|_q^q$ and observe that $\sum_{j=1}^m u_j = 1$ and $\sum_{j=1}^m v_j = 1$. By the weighted geometric mean inequality,

$$u_j^{1/p} v_j^{1/q} \leq \frac{u_j}{p} + \frac{v_j}{q}.$$

Then since $\sum_{j=1}^m u_j = 1$, $\sum_{j=1}^m v_j = 1$ and $1/p + 1/q = 1$

$$\frac{\sum_{j=1}^m |a_j b_j|}{\|\boldsymbol{a}\|_p \|\boldsymbol{b}\|_q} = \sum_{j=1}^m u_j^{1/p} v_j^{1/q} \leq \sum_{j=1}^m \left(\frac{u_j}{p} + \frac{v_j}{q} \right) = 1$$

which is (A.18). ■

Proof of Minkowski's Inequality. Se $q = p/(p-1)$ so that $1/p + 1/q = 1$. Using the triangle inequality for real numbers and two applications of Hölder's inequality

$$\begin{aligned}
\|\boldsymbol{a} + \boldsymbol{b}\|_p^p &= \sum_{j=1}^m |a_j + b_j|^p \\
&= \sum_{j=1}^m |a_j + b_j| \, |a_j + b_j|^{p-1} \\
&\leq \sum_{j=1}^m |a_j| \, |a_j + b_j|^{p-1} + \sum_{j=1}^m |b_j| \, |a_j + b_j|^{p-1} \\
&\leq \|\boldsymbol{a}\|_p \left(\sum_{j=1}^m |a_j + b_j|^{(p-1)q} \right)^{1/q} + \|\boldsymbol{b}\|_p \left(\sum_{j=1}^m |a_j + b_j|^{(p-1)q} \right)^{1/q} \\
&= \left(\|\boldsymbol{a}\|_p + \|\boldsymbol{b}\|_p \right) \|\boldsymbol{a} + \boldsymbol{b}\|_p^{p-1}
\end{aligned}$$

Solving, we find (A.19). ■

Proof of Schwarz Inequality. Using Hölder's inequality with $p = q = 2$

$$\left| \boldsymbol{a}'\boldsymbol{b} \right| \leq \sum_{j=1}^m |a_j b_j| \leq \|\boldsymbol{a}\| \, \|\boldsymbol{b}\|$$

■

A.18 Matrix Norms

Two common norms used for matrix spaces are the **Frobenius norm** and the **spectral norm**. We can write either as $\|\boldsymbol{A}\|$, but may write $\|\boldsymbol{A}\|_F$ or $\|\boldsymbol{A}\|_2$ when we want to be specific.

The **Frobenius norm** of an $m \times k$ matrix $\boldsymbol{A}$ is the Euclidean norm applied to its elements

$$\begin{aligned}
\|\boldsymbol{A}\|_F &= \|\text{vec}\,(\boldsymbol{A})\| \\
&= \left(\text{tr}\left(\boldsymbol{A}'\boldsymbol{A} \right) \right)^{1/2} \\
&= \left(\sum_{i=1}^m \sum_{j=1}^k a_{ij}^2 \right)^{1/2}.
\end{aligned}$$

When $m \times m$ $\boldsymbol{A}$ is real symmetric then

$$\|\boldsymbol{A}\|_F = \left(\sum_{\ell=1}^{m} \lambda_\ell^2\right)^{1/2}$$

where λ_ℓ, $\ell = 1, ..., m$ are the eigenvalues of $\boldsymbol{A}$. To see this, by the spectral decomposition $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ with $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_1, ..., \lambda_m\}$, so

$$\|\boldsymbol{A}\|_F = \left(\operatorname{tr}\left(\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'\right)\right)^{1/2} = \left(\operatorname{tr}\left(\boldsymbol{\Lambda}\boldsymbol{\Lambda}\right)\right)^{1/2} = \left(\sum_{\ell=1}^{m} \lambda_\ell^2\right)^{1/2}. \tag{A.22}$$

A useful calculation is for any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, using (A.1),

$$\left\|\boldsymbol{a}\boldsymbol{b}'\right\|_F = \operatorname{tr}\left(\boldsymbol{b}\boldsymbol{a}'\boldsymbol{a}\boldsymbol{b}'\right)^{1/2} = \left(\boldsymbol{b}'\boldsymbol{b}\boldsymbol{a}'\boldsymbol{a}\right)^{1/2} = \|\boldsymbol{a}\|\,\|\boldsymbol{b}\| \tag{A.23}$$

and in particular

$$\left\|\boldsymbol{a}\boldsymbol{a}'\right\|_F = \|\boldsymbol{a}\|^2. \tag{A.24}$$

The **spectral norm** of an $m \times k$ real matrix $\boldsymbol{A}$ is its largest singular value

$$\|\boldsymbol{A}\|_2 = s_{\max}(\boldsymbol{A}) = \left(\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2}$$

where $\lambda_{\max}(\boldsymbol{B})$ denotes the largest eigenvalue of the matrix $\boldsymbol{B}$. Notice that

$$\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right) = \left\|\boldsymbol{A}'\boldsymbol{A}\right\|_2$$

so

$$\|\boldsymbol{A}\|_2 = \left\|\boldsymbol{A}'\boldsymbol{A}\right\|_2^{1/2}.$$

If $\boldsymbol{A}$ is $m \times m$ and symmetric with eigenvalues λ_j then

$$\|\boldsymbol{A}\|_2 = \max_{j \le m} |\lambda_j|.$$

The Frobenius and spectral norms are closely related. They are equivalent when applied to a matrix of rank 1, since $\left\|\boldsymbol{a}\boldsymbol{b}'\right\|_2 = \|\boldsymbol{a}\|\,\|\boldsymbol{b}\| = \left\|\boldsymbol{a}\boldsymbol{b}'\right\|_F$. In general, for $m \times k$ matrix $\boldsymbol{A}$ with rank r

$$\|\boldsymbol{A}\|_2 = \left(\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} \le \left(\sum_{j=1}^{k} \lambda_j\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} = \|\boldsymbol{A}\|_F.$$

Since $\boldsymbol{A}'\boldsymbol{A}$ also has rank at most r, it has at most r non-zero eigenvalues, and hence

$$\|\boldsymbol{A}\|_F = \left(\sum_{j=1}^{k} \lambda_j\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} = \left(\sum_{j=1}^{r} \lambda_j\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} \le \left(r\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} = \sqrt{r}\,\|\boldsymbol{A}\|_2.$$

Given any vector norm $\|\boldsymbol{a}\|$ the **induced matrix norm** is defined as

$$\|\boldsymbol{A}\| = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}\| = \sup_{\boldsymbol{x} \ne 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

To see that this is a norm we need to check that it satisfies the triangle inequality. Indeed

$$\|\boldsymbol{A} + \boldsymbol{B}\| = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{x}\| \le \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}\| + \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{B}\boldsymbol{x}\| = \|\boldsymbol{A}\| + \|\boldsymbol{B}\|.$$

For any vector $\boldsymbol{x}$, by the definition of the induced norm

$$\|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}\| \|\boldsymbol{x}\|$$

a property which is called consistent norms.

Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be conformable and $\|\boldsymbol{A}\|$ an induced matrix norm. Then using the property of consistent norms

$$\|\boldsymbol{A}\boldsymbol{B}\| = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{B}\boldsymbol{x}\| \leq \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\| \|\boldsymbol{B}\boldsymbol{x}\| = \|\boldsymbol{A}\| \|\boldsymbol{B}\|.$$

A matrix norm which satisfies this property is called a **sub-multiplicative norm**, and is a matrix form of the Schwarz inequality.

Of particular interest, the matrix norm induced by the Euclidean vector norm is the spectral norm. Indeed,

$$\sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}\|^2 = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{x} = \lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right) = \|\boldsymbol{A}\|_2^2.$$

It follows that the spectral norm is consistent with the Euclidean norm, and is sub-multiplicative.

A.19 Matrix Inequalities

Schwarz Matrix Inequality: For any $m \times k$ and $k \times m$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and either the Frobenius or spectral norm,

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\| \|\boldsymbol{B}\|. \tag{A.25}$$

Triangle Inequality: For any $m \times k$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and either the Frobenius or spectral norm,

$$\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|. \tag{A.26}$$

Trace Inequality. For any $m \times m$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that $\boldsymbol{A}$ is symmetric and $\boldsymbol{B} \geq 0$

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \leq \|\boldsymbol{A}\|_2 \operatorname{tr}(\boldsymbol{B}). \tag{A.27}$$

Quadratic Inequality. For any $m \times 1$ $\boldsymbol{b}$ and $m \times m$ symmetric matrix $\boldsymbol{A}$

$$\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b} \leq \|\boldsymbol{A}\|_2 \boldsymbol{b}'\boldsymbol{b} \tag{A.28}$$

Strong Schwarz Matrix Inequality. For any conformable matrices $\boldsymbol{A}$ and $\boldsymbol{B}$

$$\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F. \tag{A.29}$$

Norm Equivalence. For any $m \times k$ matrix $\boldsymbol{A}$ of rank r

$$\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F \leq \sqrt{r}\,\|\boldsymbol{A}\|_2. \tag{A.30}$$

Eigenvalue Product Inequality. For any $m \times m$ real symmetric matrices $\boldsymbol{A} \geq 0$ and $\boldsymbol{B} \geq 0$, the eigenvalues $\lambda_\ell(\boldsymbol{A}\boldsymbol{B})$ are real and satisfy

$$\lambda_{\min}(\boldsymbol{A})\lambda_{\min}(\boldsymbol{B}) \leq \lambda_\ell(\boldsymbol{A}\boldsymbol{B}) \leq \lambda_{\max}(\boldsymbol{A})\lambda_{\max}(\boldsymbol{B}) \tag{A.31}$$

(Zhang and Zhang, 2006, Corollary 11)

Proof of Schwarz Matrix Inequality: The inequality holds for the spectral norm since it is an induced norm. Now consider the Frobenius norm. Partition $\boldsymbol{A}' = [\boldsymbol{a}_1, ..., \boldsymbol{a}_n]$ and $\boldsymbol{B} = [\boldsymbol{b}_1, ..., \boldsymbol{b}_n]$.

Then by partitioned matrix multiplication, the definition of the Frobenius norm and the Schwarz inequality for vectors

$$\begin{aligned}
\|\boldsymbol{AB}\|_F &= \left\| \begin{array}{ccc} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots \\ \vdots & \vdots & \ddots \end{array} \right\|_F \\
&\leq \left\| \begin{array}{ccc} \|\boldsymbol{a}_1\|\|\boldsymbol{b}_1\| & \|\boldsymbol{a}_1\|\|\boldsymbol{b}_2\| & \cdots \\ \|\boldsymbol{a}_2\|\|\boldsymbol{b}_1\| & \|\boldsymbol{a}_2\|\|\boldsymbol{b}_2\| & \cdots \\ \vdots & \vdots & \ddots \end{array} \right\|_F \\
&= \left(\sum_{i=1}^m \sum_{j=1}^m \|\boldsymbol{a}_i\|^2 \|\boldsymbol{b}_j\|^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^m \|\boldsymbol{a}_i\|^2 \right)^{1/2} \left(\sum_{i=1}^m \|\boldsymbol{b}_i\|^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^k \sum_{j=1}^m \boldsymbol{a}_{ji}^2 \right)^{1/2} \left(\sum_{i=1}^m \sum_{j=1}^k \|\boldsymbol{b}_{ji}\|^2 \right)^{1/2} \\
&= \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F
\end{aligned}$$

■

Proof of Triangle Inequality: The inequality holds for the spectral norm since it is an induced norm. Now consider the Frobenius norm. Let $\boldsymbol{a} = \operatorname{vec}(\boldsymbol{A})$ and $\boldsymbol{b} = \operatorname{vec}(\boldsymbol{B})$. Then by the definition of the Frobenius norm and the Schwarz Inequality for vectors

$$\begin{aligned}
\|\boldsymbol{A}+\boldsymbol{B}\|_F &= \|\operatorname{vec}(\boldsymbol{A}+\boldsymbol{B})\|_F \\
&= \|\boldsymbol{a}+\boldsymbol{b}\| \\
&\leq \|\boldsymbol{a}\| + \|\boldsymbol{b}\| \\
&= \|\boldsymbol{A}\|_F + \|\boldsymbol{B}\|_F
\end{aligned}$$

■

Proof of Trace Inequality. By the spectral decomposition for symmetric matices, $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{\Lambda}$ has the eigenvalues λ_j of $\boldsymbol{A}$ on the diagonal and $\boldsymbol{H}$ is orthonormal. Define $\boldsymbol{C} = \boldsymbol{H}'\boldsymbol{B}\boldsymbol{H}$ which has non-negative diagonal elements C_{jj} since $\boldsymbol{B}$ is positive semi-definite. Then

$$\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{\Lambda C}) = \sum_{j=1}^m \lambda_j C_{jj} \leq \max_j |\lambda_j| \sum_{j=1}^m C_{jj} = \|\boldsymbol{A}\|_2 \operatorname{tr}(\boldsymbol{C})$$

where the inequality uses the fact that $C_{jj} \geq 0$. But note that

$$\operatorname{tr}(\boldsymbol{C}) = \operatorname{tr}\left(\boldsymbol{H}'\boldsymbol{B}\boldsymbol{H}\right) = \operatorname{tr}\left(\boldsymbol{H}\boldsymbol{H}'\boldsymbol{B}\right) = \operatorname{tr}(\boldsymbol{B})$$

since $\boldsymbol{H}$ is orthonormal. Thus $\operatorname{tr}(\boldsymbol{AB}) \leq \|\boldsymbol{A}\|_2 \operatorname{tr}(\boldsymbol{B})$ as stated. ■

Proof of Quadratic Inequality: In the Trace Inequality set $\boldsymbol{B} = \boldsymbol{b}\boldsymbol{b}'$ and note $\operatorname{tr}(\boldsymbol{AB}) = \boldsymbol{b}'\boldsymbol{A}\boldsymbol{b}$ and $\operatorname{tr}(\boldsymbol{B}) = \boldsymbol{b}'\boldsymbol{b}$. ■

Proof of Strong Schwarz Matrix Inequality. By the definition of the Frobenius norm, the property of the trace, the Trace Inequality (noting that both $\boldsymbol{A}'\boldsymbol{A}$ and $\boldsymbol{B}\boldsymbol{B}'$ are symmetric and

positive semi-definite), and the Schwarz matrix inequality

$$\begin{aligned}\|\boldsymbol{AB}\|_F &= \left(\operatorname{tr}\left(\boldsymbol{B}'\boldsymbol{A}'\boldsymbol{AB}\right)\right)^{1/2} \\ &= \left(\operatorname{tr}\left(\boldsymbol{A}'\boldsymbol{ABB}'\right)\right)^{1/2} \\ &\leq \left(\left\|\boldsymbol{A}'\boldsymbol{A}\right\|_2 \operatorname{tr}\left(\boldsymbol{BB}'\right)\right)^{1/2} \\ &= \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F .\end{aligned}$$

■

Appendix B

Probability Inequalities

The following bounds are used frequently in econometric theory, predominantly in asymptotic analysis.

Monotone Probability Inequality. For any events A and B such that $A \subset B$,

$$\Pr(A) \leq \Pr(B). \tag{B.1}$$

Union Equality. For any events A and B,

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B). \tag{B.2}$$

Boole's Inequality (Union Bound). For any events A and B,

$$\Pr(A \cup B) \leq \Pr(A) + \Pr(B). \tag{B.3}$$

Bonferroni's Inequality. For any events A and B,

$$\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1. \tag{B.4}$$

Jensen's Inequality. If $g(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is convex, then for any random vector $\boldsymbol{x}$ for which $\mathbb{E}\|\boldsymbol{x}\| < \infty$ and $\mathbb{E}|g(\boldsymbol{x})| < \infty$,

$$g(\mathbb{E}(\boldsymbol{x})) \leq \mathbb{E}(g(\boldsymbol{x})). \tag{B.5}$$

If $g(\cdot)$ concave, then the inequality is reversed.

Conditional Jensen's Inequality. If $g(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is convex, then for any random vectors $(\boldsymbol{y}, \boldsymbol{x})$ for which $\mathbb{E}\|\boldsymbol{y}\| < \infty$ and $\mathbb{E}\|g(\boldsymbol{y})\| < \infty$,

$$g(\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})) \leq \mathbb{E}(g(\boldsymbol{y}) \mid \boldsymbol{x}). \tag{B.6}$$

If $g(\cdot)$ concave, then the inequality is reversed.

Conditional Expectation Inequality. For any $r \geq 1$ such that $\mathbb{E}|y|^r < \infty$, then

$$\mathbb{E}(|\mathbb{E}(y \mid \boldsymbol{x})|^r) \leq \mathbb{E}(|y|^r) < \infty. \tag{B.7}$$

Expectation Inequality. For any random matrix $\boldsymbol{Y}$ for which $\mathbb{E}\|\boldsymbol{Y}\| < \infty$,

$$\|\mathbb{E}(\boldsymbol{Y})\| \leq \mathbb{E}\|\boldsymbol{Y}\|. \tag{B.8}$$

Hölder's Inequality. If $p > 1$ and $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then for any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\mathbb{E}\left\|\boldsymbol{X}'\boldsymbol{Y}\right\| \leq \left(\mathbb{E}\left(\|\boldsymbol{X}\|^p\right)\right)^{1/p}\left(\mathbb{E}\left(\|\boldsymbol{Y}\|^q\right)\right)^{1/q}. \tag{B.9}$$

Cauchy-Schwarz Inequality. For any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\mathbb{E}\left\|\boldsymbol{X}'\boldsymbol{Y}\right\| \leq \left(\mathbb{E}\left(\|\boldsymbol{X}\|^2\right)\right)^{1/2}\left(\mathbb{E}\left(\|\boldsymbol{Y}\|^2\right)\right)^{1/2}. \tag{B.10}$$

Matrix Cauchy-Schwarz Inequality. Tripathi (1999). For any random $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^\ell$,

$$\mathbb{E}\left(\boldsymbol{y}\boldsymbol{x}'\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-}\mathbb{E}\left(\boldsymbol{x}\boldsymbol{y}'\right) \leq \mathbb{E}\left(\boldsymbol{y}\boldsymbol{y}'\right) \tag{B.11}$$

Minkowski's Inequality. For any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\left(\mathbb{E}\left(\|\boldsymbol{X}+\boldsymbol{Y}\|^p\right)\right)^{1/p} \leq \left(\mathbb{E}\left(\|\boldsymbol{X}\|^p\right)\right)^{1/p} + \left(\mathbb{E}\left(\|\boldsymbol{Y}\|^p\right)\right)^{1/p} \tag{B.12}$$

Liapunov's Inequality. For any random $m \times n$ matrix $\boldsymbol{X}$ and $1 \leq r \leq p$,

$$\left(\mathbb{E}\left(\|\boldsymbol{X}\|^r\right)\right)^{1/r} \leq \left(\mathbb{E}\left(\|\boldsymbol{X}\|^p\right)\right)^{1/p} \tag{B.13}$$

Markov's Inequality (standard form). For any random vector $\boldsymbol{x}$ and non-negative function $g(\boldsymbol{x}) \geq 0$,

$$\Pr(g(\boldsymbol{x}) > \alpha) \leq \alpha^{-1}\mathbb{E}\left(g(\boldsymbol{x})\right). \tag{B.14}$$

Markov's Inequality (strong form). For any random vector $\boldsymbol{x}$ and non-negative function $g(\boldsymbol{x}) \geq 0$,

$$\Pr(g(\boldsymbol{x}) > \alpha) \leq \alpha^{-1}\mathbb{E}\left(g\left(\boldsymbol{x}\right)1\left(g(\boldsymbol{x}) > \alpha\right)\right). \tag{B.15}$$

Chebyshev's Inequality. For any random variable x,

$$\Pr(|x - \mathbb{E}x| > \alpha) \leq \frac{\operatorname{var}(x)}{\alpha^2}. \tag{B.16}$$

Proof of Monotone Probability Inequality. Since $A \subset B$ then $B = A \cup \{B \cap A^c\}$ where A^c is the complement of A. The sets A and $\{B \cap A^c\}$ are disjoint. Thus

$$\Pr(B) = \Pr(A \cup \{B \cap A^c\}) = \Pr(A) + \Pr(B \cap A^c) \geq \Pr(A)$$

since probabilities are non-negative. Thus $\Pr(A) \leq \Pr(B)$ as claimed. ■

Proof of Union Equality. $\{A \cup B\} = A \cup \{B \cap A^c\}$ where A and $\{B \cap A^c\}$ are disjoint. Also $B = \{B \cap A\} \cup \{B \cap A^c\}$ where $\{B \cap A\}$ and $\{B \cap A^c\}$ are disjoint. These two relationships imply

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B \cap A^c)\\ \Pr(B) &= \Pr(B \cap A) + \Pr(B \cap A^c).\end{aligned}$$

Substracting,

$$\Pr(A \cup B) - \Pr(B) = \Pr(A) - \Pr(B \cap A)$$

which is (B.2) upon rearrangement. ■

Proof of Boole's Inequality. From the Union Equality and $\Pr(A \cap B) \geq 0$,

$$\begin{aligned} \Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &\leq \Pr(A) + \Pr(B) \end{aligned}$$

as claimed. ■

Proof of Bonferroni's Inequality. Rearranging the Union Equality and using $\Pr(A \cup B) \leq 1$

$$\begin{aligned} \Pr(A \cap B) &= \Pr(A) + \Pr(B) - \Pr(A \cup B) \\ &\geq \Pr(A) + \Pr(B) - 1 \end{aligned}$$

which is (B.4). ■

Proof of Jensen's Inequality. Since $g(\boldsymbol{u})$ is convex, at any point $\boldsymbol{u}$ there is a nonempty set of subderivatives (linear surfaces touching $g(\boldsymbol{u})$ at $\boldsymbol{u}$ but lying below $g(\boldsymbol{u})$ for all $\boldsymbol{u}$). Let $a + \boldsymbol{b}'\boldsymbol{u}$ be a subderivative of $g(\boldsymbol{u})$ at $\boldsymbol{u} = \mathbb{E}(\boldsymbol{x})$. Then for all $\boldsymbol{u}$, $g(\boldsymbol{u}) \geq a + \boldsymbol{b}'\boldsymbol{u}$ yet $g(\mathbb{E}(\boldsymbol{x})) = a + \boldsymbol{b}'\mathbb{E}(\boldsymbol{x})$. Applying expectations, $\mathbb{E}(g(\boldsymbol{x})) \geq a + \boldsymbol{b}'\mathbb{E}(\boldsymbol{x}) = g(\mathbb{E}(\boldsymbol{x}))$, as stated. ■

Proof of Conditional Jensen's Inequality. The same as the proof of Jensen's Inequality, but using conditional expectations. The conditional expectations exist since $\mathbb{E}\|\boldsymbol{y}\| < \infty$ and $\mathbb{E}\|g(\boldsymbol{y})\| < \infty$. ■

Proof of Conditional Expectation Inequality. As the function $|u|^r$ is convex for $r \geq 1$, the Conditional Jensen's inequality implies

$$|\mathbb{E}(y \mid \boldsymbol{x})|^r \leq \mathbb{E}(|y|^r \mid \boldsymbol{x}).$$

Taking unconditional expectations and the law of iterated expectations, we obtain

$$\mathbb{E}(|\mathbb{E}(y \mid \boldsymbol{x})|^r) \leq \mathbb{E}(\mathbb{E}(|y|^r \mid \boldsymbol{x})) = \mathbb{E}(|y|^r) < \infty$$

as required. ■

Proof of Expectation Inequality. By the Triangle inequality, for $\lambda \in [0, 1]$,

$$\|\lambda \boldsymbol{U}_1 + (1-\lambda)\boldsymbol{U}_2\| \leq \lambda \|\boldsymbol{U}_1\| + (1-\lambda)\|\boldsymbol{U}_2\|$$

which shows that the matrix norm $g(\boldsymbol{U}) = \|\boldsymbol{U}\|$ is convex. Applying Jensen's Inequality (B.5) we find (B.8). ■

Proof of Hölder's Inequality. Since $\frac{1}{p} + \frac{1}{q} = 1$ an application of the discrete Jensen's Inequality (A.13) shows that for any real a and b

$$\exp\left[\frac{1}{p}a + \frac{1}{q}b\right] \leq \frac{1}{p}\exp(a) + \frac{1}{q}\exp(b).$$

Setting $u = \exp(a)$ and $v = \exp(b)$ this implies

$$u^{1/p}v^{1/q} \leq \frac{u}{p} + \frac{v}{q}$$

and this inequality holds for any $u > 0$ and $v > 0$.

Set $u = \|\boldsymbol{X}\|^p / \mathbb{E}(\|\boldsymbol{X}\|^p)$ and $v = \|\boldsymbol{Y}\|^q / \mathbb{E}(\|\boldsymbol{Y}\|^q)$. Note that $\mathbb{E}(u) = \mathbb{E}(v) = 1$. By the matrix Schwarz Inequality (A.25), $\|\boldsymbol{X}'\boldsymbol{Y}\| \leq \|\boldsymbol{X}\| \|\boldsymbol{Y}\|$. Thus

$$\begin{aligned}
\frac{\mathbb{E}\|\boldsymbol{X}'\boldsymbol{Y}\|}{(\mathbb{E}(\|\boldsymbol{X}\|^p))^{1/p}(\mathbb{E}(\|\boldsymbol{Y}\|^q))^{1/q}} &\leq \frac{\mathbb{E}(\|\boldsymbol{X}\|\|\boldsymbol{Y}\|)}{(\mathbb{E}(\|\boldsymbol{X}\|^p))^{1/p}(\mathbb{E}(\|\boldsymbol{Y}\|^q))^{1/q}} \\
&= \mathbb{E}\left(u^{1/p}v^{1/q}\right) \\
&\leq \mathbb{E}\left(\frac{u}{p} + \frac{v}{q}\right) \\
&= \frac{1}{p} + \frac{1}{q} \\
&= 1,
\end{aligned}$$

which is (B.9). ■

Proof of Cauchy-Schwarz Inequality. Special case of Hölder's with $p = q = 2$.

Proof of Matrix Cauchy-Schwarz Inequality. Define $e = \boldsymbol{y} - (\mathbb{E}(\boldsymbol{y}\boldsymbol{x}'))(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-}\boldsymbol{x}$. Note that $\mathbb{E}(\boldsymbol{e}\boldsymbol{e}') \geq 0$ is positive semi-definite. We can calculate that

$$\mathbb{E}(\boldsymbol{e}\boldsymbol{e}') = \mathbb{E}(\boldsymbol{y}\boldsymbol{y}') - (\mathbb{E}(\boldsymbol{y}\boldsymbol{x}'))(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-}\mathbb{E}(\boldsymbol{x}\boldsymbol{y}').$$

Since the left-hand-side is positive semi-definite, so is the right-hand-side, which means $\mathbb{E}(\boldsymbol{y}\boldsymbol{y}') \geq (\mathbb{E}(\boldsymbol{y}\boldsymbol{x}'))(\mathbb{E}(\boldsymbol{x}\boldsymbol{x}'))^{-}\mathbb{E}(\boldsymbol{x}\boldsymbol{y}')$ as stated. ■

Proof of Liapunov's Inequality. The function $g(u) = u^{p/r}$ is convex for $u > 0$ since $p \geq r$. Set $u = \|\boldsymbol{X}\|^r$. By Jensen's inequality, $g(\mathbb{E}(u)) \leq \mathbb{E}(g(u))$ or

$$(\mathbb{E}(\|\boldsymbol{X}\|^r))^{p/r} \leq \mathbb{E}\left((\|\boldsymbol{X}\|^r)^{p/r}\right) = \mathbb{E}(\|\boldsymbol{X}\|^p).$$

Raising both sides to the power $1/p$ yields $(\mathbb{E}(\|\boldsymbol{X}\|^r))^{1/r} \leq (\mathbb{E}(\|\boldsymbol{X}\|^p))^{1/p}$ as claimed. ■

Proof of Minkowski's Inequality. Note that by rewriting, using the triangle inequality (A.26), and then Hölder's Inequality to the two expectations

$$\begin{aligned}
\mathbb{E}(\|\boldsymbol{X} + \boldsymbol{Y}\|^p) &= \mathbb{E}\left(\|\boldsymbol{X} + \boldsymbol{Y}\| \|\boldsymbol{X} + \boldsymbol{Y}\|^{p-1}\right) \\
&\leq \mathbb{E}\left(\|\boldsymbol{X}\| \|\boldsymbol{X} + \boldsymbol{Y}\|^{p-1}\right) + \mathbb{E}\left(\|\boldsymbol{Y}\| \|\boldsymbol{X} + \boldsymbol{Y}\|^{p-1}\right) \\
&\leq (\mathbb{E}(\|\boldsymbol{X}\|^p))^{1/p}\,\mathbb{E}\left(\left(\|\boldsymbol{X} + \boldsymbol{Y}\|^{q(p-1)}\right)^{1/q}\right) \\
&+ (\mathbb{E}(\|\boldsymbol{Y}\|^p))^{1/p}\,\mathbb{E}\left(\left(\|\boldsymbol{X} + \boldsymbol{Y}\|^{q(p-1)}\right)^{1/q}\right) \\
&= \left((\mathbb{E}(\|\boldsymbol{X}\|^p))^{1/p} + (\mathbb{E}(\|\boldsymbol{Y}\|^p))^{1/p}\right)\mathbb{E}\left((\|\boldsymbol{X} + \boldsymbol{Y}\|^p)^{(p-1)/p}\right)
\end{aligned}$$

where the second equality picks q to satisfy $1/p + 1/q = 1$, and the final equality uses this fact to make the substitution $q = p/(p-1)$ and then collects terms. Dividing both sides by $\mathbb{E}\left((\|\boldsymbol{X} + \boldsymbol{Y}\|^p)^{(p-1)/p}\right)$, we obtain (B.12). ■

Proof of Markov's Inequality. Let F denote the distribution function of $\boldsymbol{x}$. Then

$$\begin{aligned}\Pr\left(g(\boldsymbol{x}) \geq \alpha\right) &= \int_{\{g(\boldsymbol{u})\geq\alpha\}} dF(\boldsymbol{u}) \\ &\leq \int_{\{g(\boldsymbol{u})\geq\alpha\}} \frac{g(\boldsymbol{u})}{\alpha} dF(\boldsymbol{u}) \\ &= \alpha^{-1} \int 1\left(g(\boldsymbol{u}) > \alpha\right) g(\boldsymbol{u}) dF(\boldsymbol{u}) \\ &= \alpha^{-1}\mathbb{E}\left(g\left(\boldsymbol{x}\right) 1\left(g(\boldsymbol{x}) > \alpha\right)\right)\end{aligned}$$

the inequality using the region of integration $\{g(\boldsymbol{u}) > \alpha\}$. This establishes the strong form (B.15). Since $1\left(g(\boldsymbol{x}) > \alpha\right) \leq 1$, the final expression is less than $\alpha^{-1}\mathbb{E}\left(g(\boldsymbol{x})\right)$, establishing the standard form (B.14). ■

Proof of Chebyshev's Inequality. Define $y = (x - \mathbb{E}x)^2$ and note that $\mathbb{E}\left(y\right) = \operatorname{var}\left(x\right)$. The events $\{|x - \mathbb{E}x| > \alpha\}$ and $\left\{y > \alpha^2\right\}$ are equal, so by an application Markov's inequality we find

$$\Pr(|x - \mathbb{E}x| > \alpha) = \Pr(y > \alpha^2) \leq \alpha^{-2}\mathbb{E}\left(y\right) = \alpha^{-2}\operatorname{var}\left(x\right)$$

as stated. ■

Bibliography

[1] Abadir, Karim M. and Jan R. Magnus (2005): *Matrix Algebra*, Cambridge University Press.

[2] Acemoglu, Daron, Simon Johnson, James A. Robinson (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91, 1369-1401.

[3] Acemoglu, Daron, Simon Johnson, James A. Robinson (2012): "The Colonial Origins of Comparative Development: An Empirical Investigation: Reply," *American Economic Review*, 102, 3077–3110.

[4] Aitken, A.C. (1935): "On least squares and linear combinations of observations," *Proceedings of the Royal Statistical Society*, 55, 42-48.

[5] Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle." In B. Petroc and F. Csake, eds., *Second International Symposium on Information Theory*.

[6] Anderson, T.W. and H. Rubin (1949): "Estimation of the parameters of a single equation in a complete system of stochastic equations," *The Annals of Mathematical Statistics*, 20, 46-63.

[7] Andrews, Donald W. K. (1988): "Laws of large numbers for dependent non-identically distributed random variables," *Econometric Theory,* 4, 458-467.

[8] Andrews, Donald W. K. (1991), "Asymptotic normality of series estimators for nonparameric and semiparametric regression models," *Econometrica*, 59, 307-345.

[9] Andrews, Donald W. K. (1993), "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821-8516.

[10] Andrews, Donald W. K. and Moshe Buchinsky: (2000): "A three-step method for choosing the number of bootstrap replications," *Econometrica*, 68, 23-51.

[11] Andrews, Donald W. K. and Werner Ploberger (1994): "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica,* 62, 1383-1414.

[12] Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996): "Identification of causal effects using instrumental variables," *Journal of the American Statistical Association*, 55, 650-659.

[13] Angrist, Joshua D. and Alan B. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?" *Quarterly Journal of Economics*, 91, 444-455.

[14] Ash, Robert B. (1972): *Real Analysis and Probability*, Academic Press.

[15] Barro, Robert J. (1977): "Unanticipated money growth and unemployment in the United States," *American Economic Review*, 67, 101–115

[16] Basmann, R. L. (1957): "A generalized classical method of linear estimation of coefficients in a structural equation," *Econometrica,* 25, 77-83.

[17] Basmann, R. L. (1960): "On finite sample distributions of generalized classical linear identifiability test statistics," *Journal of the American Statistical Association*, 55, 650-659.

[18] Baum, Christopher F, Mark E. Schaffer, and Steven Stillman (2003): "Instrumental variables and GMM: Estimation and testing," *The Stata Journal*, 3, 1-31.

[19] Bekker, P.A. (1994): "Alternative approximations to the distributions of instrumental variable estimators, *Econometrica,* 62, 657-681.

[20] Billingsley, Patrick (1968): *Convergence of Probability Measures.* New York: Wiley.

[21] Billingsley, Patrick (1995): *Probability and Measure,* 3rd Edition, New York: Wiley.

[22] Bose, A. (1988): "Edgeworth correction by bootstrap in autoregressions," *Annals of Statistics*, 16, 1709-1722.

[23] Box, George E. P. and Dennis R. Cox, (1964). "An analysis of transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211-252.

[24] Breusch, T.S. and A.R. Pagan (1979): "The Lagrange multiplier test and its application to model specification in econometrics," *Review of Economic Studies*, 47, 239-253.

[25] Brown, B. W. and Whitney K. Newey (2002): "GMM, efficient bootstrapping, and improved inference ," *Journal of Business and Economic Statistics.*

[26] Card, David (1995): "Using geographic variation in college proximity to estimate the return to schooling," in *Aspects of Labor Market Behavior: Essays in Honour of John Vanderkamp*, L.N. Christofides, E.K. Grant, and R. Swidinsky, editors. Toronto: University of Toronto Press.

[27] Carlstein, E. (1986): "The use of subseries methods for estimating the variance of a general statistic from a stationary time series," *Annals of Statistics*, 14, 1171-1179.

[28] Casella, George and Roger L. Berger (2002): *Statistical Inference*, 2nd Edition, Duxbury Press.

[29] Chamberlain, Gary (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.

[30] Chow, G.C. (1960): "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, 28, 591-603.

[31] Cragg, John G. (1992): "Quasi-Aitken Estimation for Heterskedasticity of Unknown Form" *Journal of Econometrics*, 54, 179-201.

[32] Cragg, John G. and Stephen G. Donald (1993): "Testing identifiability and specification in instrumental variable models," *Econometric Theory*, 9, 222-240.

[33] Davidson, James (1994): *Stochastic Limit Theory: An Introduction for Econometricians.* Oxford: Oxford University Press.

[34] Davison, A.C. and D.V. Hinkley (1997): *Bootstrap Methods and their Application.* Cambridge University Press.

[35] De Luca, Giuseppe, Jan R. Magnus, and Franco Peracchi (2017): "Balanced variable addition in linear models" *Journal of Economic Surveys*, 31.

[36] Dickey, D.A. and W.A. Fuller (1979): "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American Statistical Association*, 74, 427-431.

[37] Donald Stephen G. and Whitney K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161-1191.

[38] Duflo, Esther, Pascaline Dupas, and Michael Kremer (2011): "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review*, 101, 1739-1774.

[39] Dufour, Jean-Marie (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365-1387.

[40] Durbin, James (1954): "Errors in variables," *Review of the International Statistical Institute*, 22, 23-32.

[41] Efron, Bradley (1979): "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1-26.

[42] Efron, Bradley (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

[43] Efron, Bradley and R.J. Tibshirani (1993): *An Introduction to the Bootstrap*, New York: Chapman-Hall.

[44] Eichenbaum, Martin S., Lars Peter Hansen, and Kenneth J. Singleton (1988): "A time series analysis of representative agent models of consumption and leisure choice," *The Quarterly Journal of Economics*, 103, 51-78.

[45] Eicker, F. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34, 447-456.

[46] Engle, Robert F. and Clive W. J. Granger (1987): "Co-integration and error correction: Representation, estimation and testing," *Econometrica*, 55, 251-276.

[47] Frisch, Ragnar (1933): "Editorial," *Econometrica*, 1, 1-4.

[48] Frisch, Ragnar and F. Waugh (1933): "Partial time regressions as compared with individual trends," *Econometrica*, 1, 387-401.

[49] Gallant, A. Ronald and D. W. Nychka (1987): "Seminonparametric maximum likelihood estimation," *Econometrica*, 55, 363-390.

[50] Gallant, A. Ronald and Halbert White (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. New York: Basil Blackwell.

[51] Galton, Francis (1886): "Regression Towards Mediocrity in Hereditary Stature," The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

[52] Goldberger, Arthur S. (1964): *Econometric Theory*, Wiley.

[53] Goldberger, Arthur S. (1968): *Topics in Regression Analysis*, Macmillan

[54] Goldberger, Arthur S. (1991): *A Course in Econometrics*. Cambridge: Harvard University Press.

[55] Goffe, W.L., G.D. Ferrier and J. Rogers (1994): "Global optimization of statistical functions with simulated annealing," *Journal of Econometrics*, 60, 65-99.

[56] Gosset, William S. (a.k.a. "Student") (1908): "The probable error of a mean," *Biometrika*, 6, 1-25.

[57] Gauss, K. F. (1809): "Theoria motus corporum coelestium," in *Werke*, Vol. VII, 240-254.

[58] Granger, Clive W. J. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37, 424-438.

[59] Granger, Clive W. J. (1981): "Some properties of time series data and their use in econometric specification," *Journal of Econometrics*, 16, 121-130.

[60] Granger, Clive W. J. and Timo Teräsvirta (1993): *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

[61] Gregory, A. and M. Veall (1985): "On formulating Wald tests of nonlinear restrictions," *Econometrica*, 53, 1465-1468,

[62] Haavelmo, T. (1944): "The probability approach in econometrics," *Econometrica*, supplement, 12.

[63] Hall, A. R. (2000): "Covariance matrix estimation and the power of the overidentifying restrictions test," *Econometrica*, 68, 1517-1527.

[64] Hall, B. H. and R. E. Hall (1993): "The Value and Performance of U.S. Corporations" (1993) *Brookings Papers on Economic Activity*, 1-49.

[65] Hall, Peter (1992): *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

[66] Hall, Peter (1994): "Methodology and theory for the bootstrap," *Handbook of Econometrics, Vol. IV*, eds. R.F. Engle and D.L. McFadden. New York: Elsevier Science.

[67] Hall, Peter and Joel L. Horowitz (1996): "Bootstrap critical values for tests based on Generalized-Method-of-Moments estimation," *Econometrica*, 64, 891-916.

[68] Hahn, Jinyong (1996): "A note on bootstrapping generalized method of moments estimators," *Econometric Theory*, 12, 187-197.

[69] Hamilton, James D. (1994) *Time Series Analysis*.

[70] Hansen, Bruce E. (1992): "Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends," *Journal of Econometrics*, 53, 87-121.

[71] Hansen, Bruce E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64, 413-430.

[72] Hansen, Bruce E. (1999): "Threshold effects in non-dynamic panels: Estimation, testing and inference," *Journal of Econometrics*, 93, 345-368.

[73] Hansen, Bruce E. (2006): "Edgeworth expansions for the Wald and GMM statistics for nonlinear restrictions," *Econometric Theory and Practice: Frontiers of Analysis and Applied Research*, edited by Dean Corbae, Steven N. Durlauf and Bruce E. Hansen. Cambridge University Press.

[74] Hansen, Bruce E. and Seojeong Lee (2018): "Inference for iterated GMM under misspecification and clustering", working paper.

[75] Hansen, Christoper B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large, *Journal of Econometrics*, 141, 595-620.

[76] Hansen, Lars Peter (1982): "Large sample properties of generalized method of moments estimators, *Econometrica*, 50, 1029-1054.

[77] Hansen, Lars Peter, John Heaton, and A. Yaron (1996): "Finite sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics*, 14, 262-280.

[78] Hausman, J.A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251-1271.

[79] Heckman, J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153-161.

[80] Hinkley, D. V. (1977): "Jackknifing in unbalanced situations," *Technometrics*, 19, 285-292.

[81] Horn, S.D., R.A. Horn, and D.B. Duncan. (1975) "Estimating heteroscedastic variances in linear model," *Journal of the American Statistical Association*, 70, 380-385.

[82] Horowitz, Joel (2001): "The Bootstrap," *Handbook of Econometrics, Vol. 5*, J.J. Heckman and E.E. Leamer, eds., Elsevier Science, 3159-3228.

[83] Imbens, Guido W. (1997): "One step estimators for over-identified generalized method of moments models," Review of Economic Studies, 64, 359-383.

[84] Imbens, Guido W., and Joshua D. Angrist (1994): "Identification and estimation of local average treatment effects," *Econometrica*, 62, 467-476.

[85] Imbens, G.W., R.H. Spady and P. Johnson (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333-357.

[86] Jarque, C.M. and A.K. Bera (1980): "Efficient tests for normality, homoskedasticity and serial independence of regression residuals, *Economic Letters*, 6, 255-259.

[87] Johansen, S. (1988): "Statistical analysis of cointegrating vectors," *Journal of Economic Dynamics and Control*, 12, 231-254.

[88] Johansen, S. (1991): "Estimation and hypothesis testing of cointegration vectors in the presence of linear trend," *Econometrica*, 59, 1551-1580.

[89] Johansen, S. (1995): *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*, Oxford University Press.

[90] Johansen, S. and K. Juselius (1992): "Testing structural hypotheses in a multivariate cointegration analysis of the PPP and the UIP for the UK," *Journal of Econometrics*, 53, 211-244.

[91] Kilian, Lutz and Helmut Lütkepohl: (2017): *Structural Vector Autoregressive Analysis*, Cambridge University Press, forthcoming.

[92] Kitamura, Y. (2001): "Asymptotic optimality and empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661-1672.

[93] Kitamura, Y. and M. Stutzer (1997): "An information-theoretic alternative to generalized method of moments," *Econometrica*, 65, 861-874..

[94] Koenker, Roger (2005): *Quantile Regression*. Cambridge University Press.

[95] Kunsch, H.R. (1989): "The jackknife and the bootstrap for general stationary observations," *Annals of Statistics*, 17, 1217-1241.

[96] Kwiatkowski, D., P.C.B. Phillips, P. Schmidt, and Y. Shin (1992): "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?" *Journal of Econometrics*, 54, 159-178.

[97] Lafontaine, F. and K.J. White (1986): "Obtaining any Wald statistic you want," *Economics Letters*, 21, 35-40.

[98] Legendre, Adrien-Marie (1805): *Nouvelles methodes pour la determination des orbites de cometes [New Methods for the Determination of the Orbits of Comets]*, Pris: F. Didot.

[99] Lehmann, E.L. and George Casella (1998): *Theory of Point Estimation*, 2^{nd} Edition, Springer.

[100] Lehmann, E.L. and Joseph P. Romano (2005): *Testing Statistical Hypotheses*, 3^{rd} Edition, Springer.

[101] Lindeberg, Jarl Waldemar, (1922): "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, 15, 211-225.

[102] Li, Qi and Jeffrey Racine (2007) *Nonparametric Econometrics*.

[103] Lovell, M.C. (1963): "Seasonal adjustment of economic time series," *Journal of the American Statistical Association*, 58, 993-1010.

[104] MacKinnon, James G. (1990): "Critical values for cointegration," in Engle, R.F. and C.W. Granger (eds.) *Long-Run Economic Relationships: Readings in Cointegration*, Oxford, Oxford University Press.

[105] MacKinnon, James G. and Halbert White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305-325.

[106] Magnus, J. R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: John Wiley and Sons.

[107] Mankiw, N. Gregory, David Romer, and David N. Weil (1992): "A contribution to the empirics of economic growth," *The Quarterly Journal of Economics*, 107, 407-437.

[108] Mann, H.B. and A. Wald (1943). "On stochastic limit and order relationships," *The Annals of Mathematical Statistics* 14, 217–226.

[109] Muirhead, R.J. (1982): *Aspects of Multivariate Statistical Theory*. New York: Wiley.

[110] Nelder, J. and R. Mead (1965): "A simplex method for function minimization," *Computer Journal*, 7, 308-313.

[111] Nerlove, Marc (1963): "Returns to Scale in Electricity Supply," Chapter 7 of *Measurement in Economics* (C. Christ, et al, eds.). Stanford: Stanford University Press, 167-198.

[112] Newey, Whitney K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99-135.

[113] Newey, Whitney K. (1995): "Generalized method of moments specification testing," *Journal of Econometrics*, 29, 229-256.

[114] Newey, Whitney K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147-168.

[115] Newey, Whitney K. and Daniel L. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in Robert Engle and Daniel McFadden, (eds.) *Handbook of Econometrics*, vol. IV, 2111-2245, North Holland: Amsterdam.

[116] Newey, Whitney K. and Kenneth D. West (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.

[117] Owen, Art B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237-249.

[118] Owen, Art B. (2001): *Empirical Likelihood*. New York: Chapman & Hall.

[119] Pagan, Adrian (1984): "Econometric issues in the analysis of regressions with generated regressors," *International Economic Review*, 25, 221-247.

[120] Pagan, Adrian (1986): "Two stage and related estimators and their applications," *Review of Economic Studies*, 53, 517-538.

[121] Park, Joon Y. and Peter C. B. Phillips (1988): "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, 56, 1065-1083,

[122] Phillips, Peter C.B. (1983): "Exact small sample theory in the simultaneous equatios model," *Handbook of Econometrics, Volume 1*, edited by Z. Griliches and M. D. Intriligator, North-Holland.

[123] Phillips, Peter C.B. and Sam Ouliaris (1990): "Asymptotic properties of residual based tests for cointegration," *Econometrica*, 58, 165-193.

[124] Politis, D.N. and J.P. Romano (1996): "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303-1313.

[125] Potscher, B.M. (1991): "Effects of model selection on inference," *Econometric Theory*, 7, 163-185.

[126] Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300-325.

[127] Ramsey, J. B. (1969): "Tests for specification errors in classical linear least-squares regression analysis," *Journal of the Royal Statistical Society*, Series B, 31, 350-371.

[128] Rudin, W. (1987): *Real and Complex Analysis*, 3rd edition. New York: McGraw-Hill.

[129] Runge, Carl (1901): "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten," *Zeitschrift für Mathematik und Physik*, 46, 224-243.

[130] Said, S.E. and D.A. Dickey (1984): "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, 71, 599-608.

[131] Sargan, J.D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica, 2*6, 393-415.

[132] Secrist, Horace (1933): *The Triumph of Mediocrity in Business*. Evanston: Northwestern University.

[133] Shao, Jun and D. Tu (1995): *The Jackknife and Bootstrap*. NY: Springer.

[134] Shao, Jun (2003): *Mathematical Statistics*, 2nd edition, Springer.

[135] Sheather, S.J. and M.C. Jones (1991): "A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society, Series B*, 53, 683-690.

[136] Shin, Y. (1994): "A residual-based test of the null of cointegration against the alternative of no cointegration," *Econometric Theory*, 10, 91-115.

[137] Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

[138] Sims, C.A. (1972): "Money, income and causality," *American Economic Review*, 62, 540-552.

[139] Sims, C.A. (1980): "Macroeconomics and reality," *Econometrica, 48*, 1-48.

[140] Staiger, D. and James H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica,* 65, 557-586.

[141] Stock, James H. (1987): "Asymptotic properties of least squares estimators of cointegrating vectors," *Econometrica,* 55, 1035-1056.

[142] Stock, James H. (1991): "Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series," *Journal of Monetary Economics,* 28, 435-460.

[143] Stock, James H. and Jonathan H. Wright (2000): "GMM with weak identification," *Econometrica,* 68, 1055-1096.

[144] Stock, James H. and Mark W. Watson (2014): *Introduction to Econometrics*, 3^{rd} edition, Pearson.

[145] Stock, James H. and Motohiro Yogo (2005): "Testing for weak instruments in linear IV regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, eds Donald W.K. Andrews and James H. Stock, Cambridge University Press, 80-108.

[146] Stone, Marshall H. (1937): "Applications of the Theory of Boolean Rings to General Topology," *Transactions of the American Mathematical Society*, 41, 375-481.

[147] Stone, Marshall H. (1948): "The Generalized Weierstrass Approximation Theorem," *Mathematics Magazine*, 21, 167-184.

[148] Theil, Henri. (1953): "Repeated least squares applied to complete equation systems," The Hague, Central Planning Bureau, mimeo.

[149] Theil, Henri (1961): *Economic Forecasts and Policy*. Amsterdam: North Holland.

[150] Theil, Henri. (1971): *Principles of Econometrics*, New York: Wiley.

[151] Tobin, James (1958): "Estimation of relationships for limited dependent variables," *Econometrica, 2*6, 24-36.

[152] Tripathi, Gautam (1999): "A matrix extension of the Cauchy-Schwarz inequality," Economics Letters, 63, 1-3.

[153] van der Vaart, A.W. (1998): *Asymptotic Statistics*, Cambridge University Press.

[154] Wald, Abraham. (1940): "The fitting of straight lines if both variables are subject to error," *The Annals of Mathematical Statistics*, 11, 283-300

[155] Wald, Abraham. (1943): "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, 54, 426-482.

[156] Weierstrass, K. (1885): "Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen," *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, 1885.

[157] White, Halbert (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817-838.

[158] White, Halbert (1984): *Asymptotic Theory for Econometricians,* Academic Press.

[159] Wooldridge, Jeffrey M. (1995): "Score diagnostics for linear models estimated by two stage least squares," In *Advances in Econometrics and Quantitative Economics: Essays in honor of Professor C. R. Rao*, eds. G. S. Maddala, P.C.B. Phillpis, and T.N. Srinivasan, 66-87. Cambridge: Blackwell.

[160] Wooldridge, Jeffrey M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2^{nd} edition, MIT Press.

[161] Wooldridge, Jeffrey M. (2015) *Introductory Econometrics: A Modern Approach,* 6^{th} edition, Southwestern.

[162] Wu, De-Min (1973): Alternative tests of independence between stochastic regressors and disturbances," *Econometrica,* 41, 733-750.

[163] Zellner, Arnold. (1962): "An efficient method of estimating seemingly unrelated regressions, and tests for aggregation bias," *Journal of the American Statistical Association*, 57, 348-368.

[164] Zhang, Fuzhen and Qingling Zhang (2006): "Eigenvalue inequalities for matrix product," *IEEE Transactions on Automatic Control*, 51, 1506-1509.)