ECONOMETRICS

BRUCE E. HANSEN

University of Wisconsin

Department of Economics

This Revision: January 14, 2016
Comments Welcome

[1] This manuscript may be printed and reproduced for individual or instructional use, but may not be printed for commercial purposes.

Contents

Preface

This book is intended to serve as the textbook for a first-year graduate course in econometrics. It can be used as a stand-alone text, or be used as a supplement to another text.

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 (2009) and Stock and Watson (2010).

For reference, some of the basic tools of matrix algebra, probability, and statistics 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) for time-series methods, Wooldridge (2002) 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 for providing research assistance in preparing some of the empirical examples presented in the text.

As this is a manuscript in progress, some parts are quite incomplete, and there are many topics which I plan to add. In general, the earlier chapters are the most complete while the later chapters need significant work and revision.

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", *Econometrica* (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 an 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. $\hat{\beta}$ and $\tilde{\beta}$ 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 distibution 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 three major types of economic data sets: cross-sectional, time-series, and panel. They are distinguished by the dependence structure across observations.

Cross-sectional data sets have one observation per individual. Surveys 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 so the random sampling assumption is inappropriate. 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. The 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) surveyed 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. This is a modified random sampling environment.

Data Structures

- Cross-section
- Time-series
- Panel

Many contemporary econometric applications combine elements of cross-section, time-series, and panel data modeling. These include models of spatial correlation and clustering.

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 $\boldsymbol{z}_i$.)

Furthermore, if the data is randomly gathered, it is reasonable to model each observation as a random draw from the same probability distribution. In this case we say that the data are **independent and identically distributed** or **iid**. We call this 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* ***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 intellectural 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 (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 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. Never-the-less, 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.

Never-the-less, 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.

1.9 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
$\operatorname{var}(y)$	variance
$\operatorname{cov}(x, y)$	covariance
$\operatorname{var}(\boldsymbol{x})$	covariance matrix
$\operatorname{corr}(x, y)$	correlation
$\Pr$	probability
$\longrightarrow$	limit
$\xrightarrow{p}$	convergence in probability
$\xrightarrow{d}$	convergence in distribution
$\operatorname{plim}_{n\to\infty}$	probability limit
$\mathrm{N}(\mu, \sigma^2)$	normal distribution
$\mathrm{N}(0, 1)$	standard normal distribution
χ^2_k	chi-square distribution with k degrees of freedom
$\boldsymbol{I}_n$	identity matrix
$\operatorname{tr} \boldsymbol{A}$	trace
$\boldsymbol{A}'$	matrix transpose
$\boldsymbol{A}^{-1}$	matrix inverse
$\boldsymbol{A} > 0$	positive definite
$\boldsymbol{A} \geq 0$	positive semi-definite
$\lVert \boldsymbol{a} \rVert$	Euclidean norm
$\lVert \boldsymbol{A} \rVert$	matrix (Frobinius) norm
$\approx$	approximate equality
$\overset{def}{=}$	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.

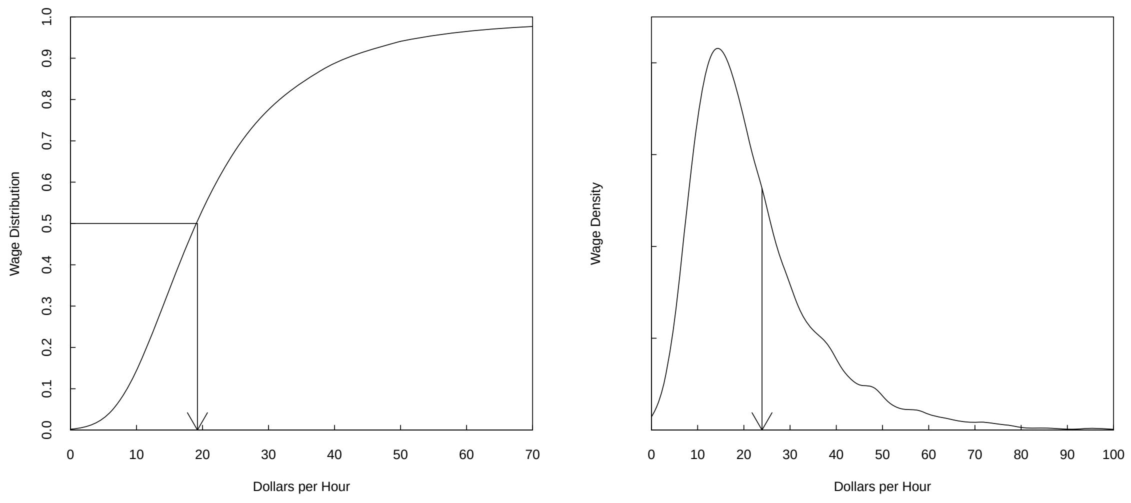


Figure 2.1: Wage Distribution and Density. All full-time U.S. workers

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

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

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 2.31. 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 the notation $\mathbb{E}y$ instead of $\mathbb{E}\left(y\right)$ 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.

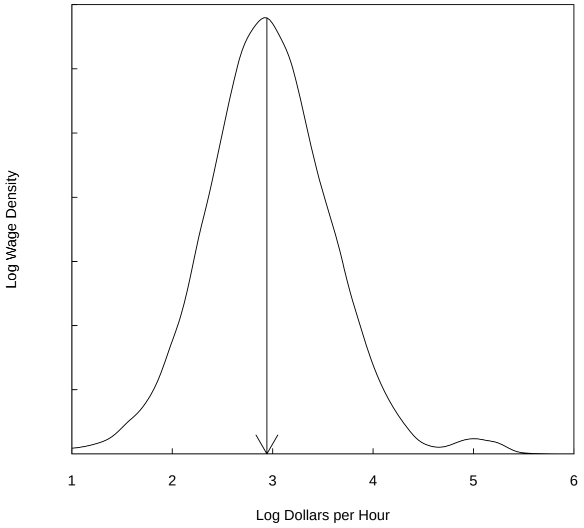


Figure 2.2: Log Wage Density

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

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

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

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

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.

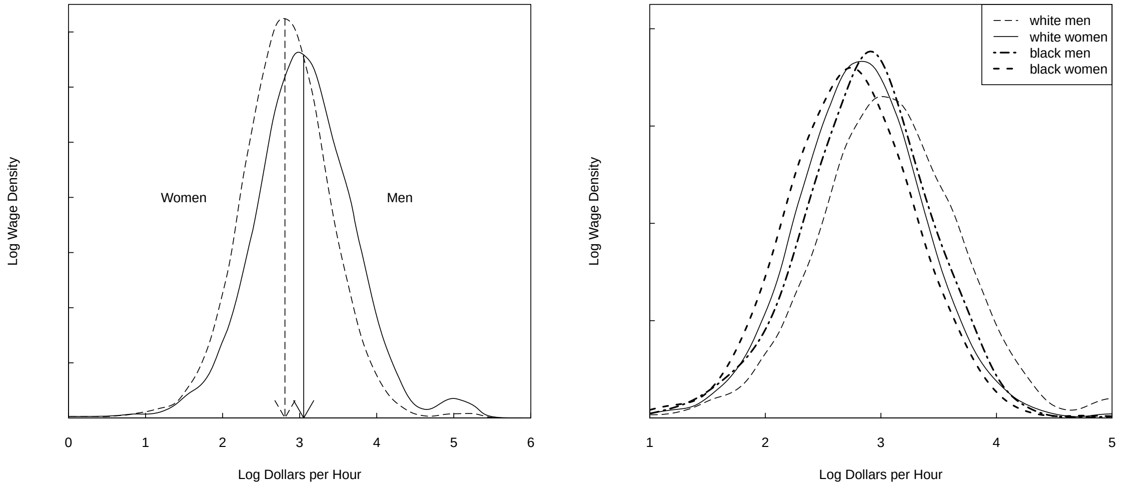
2.3 Conditional Expectation

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 $\exp\left(\mathbb{E}\left(\log w\right)\right) = \$19.11$ 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 $|c| \leq 10$.

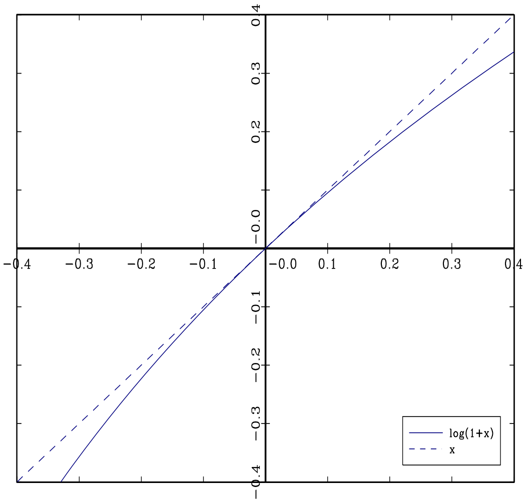


Figure 2.4: $\log(1+x)$

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 of schooling, and we will write this variable as *education*[8].

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.

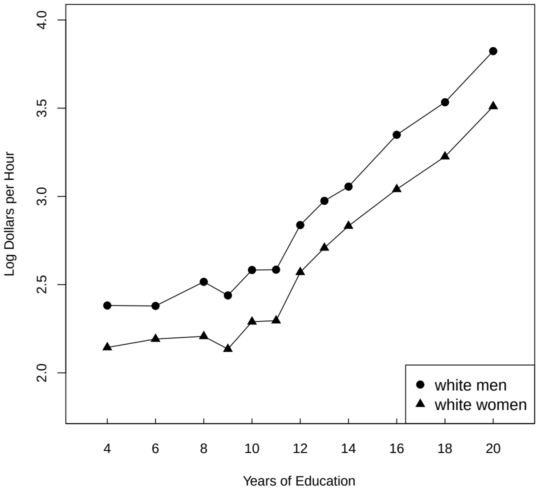


Figure 2.5: Mean Log Wage as a Function of Years of Education

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

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

The CEF $\mathbb{E}\left(y \mid \boldsymbol{x}\right)$ 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 $m\left(\boldsymbol{u}\right)=\mathbb{E}\left(y \mid \boldsymbol{x}=\boldsymbol{u}\right)$, which is a function of the argument $\boldsymbol{u}$. The expression $\mathbb{E}\left(y \mid \boldsymbol{x}=\boldsymbol{u}\right)$ 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 $\mathbb{E}\left(y \mid \boldsymbol{x}\right)$ to refer to this function. Hopefully, the use of $\mathbb{E}\left(y \mid \boldsymbol{x}\right)$ 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

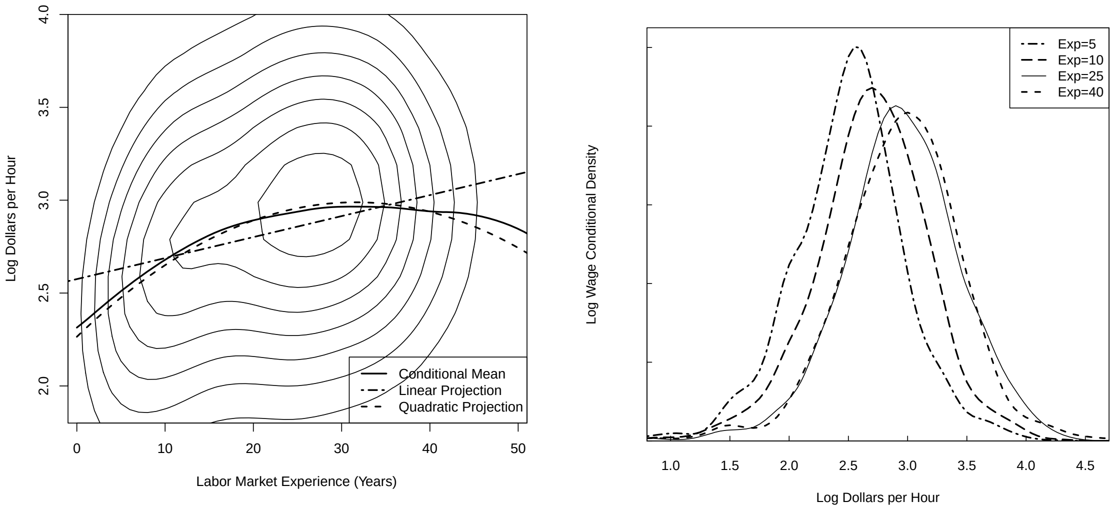
$$f_{\boldsymbol{x}}(\boldsymbol{x})=\int_{\mathbb{R}} f(y, \boldsymbol{x}) dy.$$

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

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

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)

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

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.

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

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

and when $\boldsymbol{x}$ is continuous

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

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

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

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 $\mathbb{E}\left|y\right| < \infty$ then for any random vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$,

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

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 $\mathbb{E}\left(y \mid \boldsymbol{x}_1\right)$, the expectation conditional on $\boldsymbol{x}_1$ alone. Sometimes we phrase this as: "The smaller information set wins."

As an example

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

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, $\mathbb{E}\left(\boldsymbol{x} \mid \boldsymbol{x}\right) = \boldsymbol{x}$ and $\mathbb{E}\left(g\left(\boldsymbol{x}\right) \mid \boldsymbol{x}\right) = g\left(\boldsymbol{x}\right)$ for any function $g(\cdot)$. The general property is known as the Conditioning Theorem.

Theorem 2.7.3 ***Conditioning Theorem***
If

$$\mathbb{E}\left|g\left(\boldsymbol{x}\right) y\right| < \infty \tag{2.8}$$

then

$$\mathbb{E}\left(g\left(\boldsymbol{x}\right) y \mid \boldsymbol{x}\right) = g\left(\boldsymbol{x}\right) \mathbb{E}\left(y \mid \boldsymbol{x}\right) \tag{2.9}$$

and

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

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.

The equations

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

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 $\mathbb{E}(e \mid \boldsymbol{x}) = 0$ is implied by the definition of e as the difference between y and the CEF $m(\boldsymbol{x})$. The equation $\mathbb{E}(e \mid \boldsymbol{x}) = 0$ 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

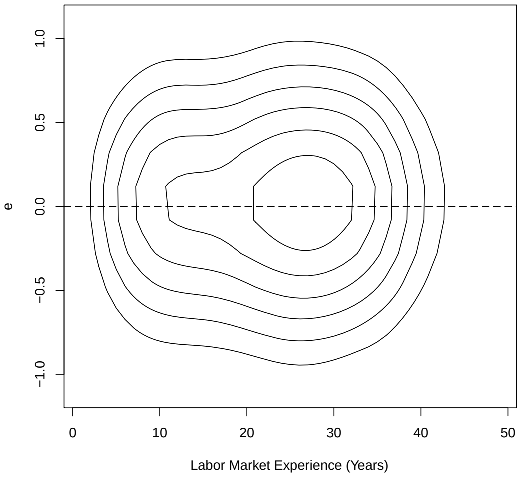


Figure 2.7: Joint density of CEF error e and *experience* for white men with *education*=12.

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 $\mathbb{E}(e \mid x) = 0$ 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 $m(\boldsymbol{x}) = \mathbb{E}(y) = \mu$, the unconditional mean of y. We can still write an equation for y in the regression format:

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

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* $\mathbb{E}y^2 < \infty$ *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

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

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* $\mathbb{E}y^2 < \infty$ *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

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

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

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

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 $\mathbb{E}y^2 < \infty$ 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 $\mathbb{E}y^2 < \infty$, *then for any predictor* $g(\boldsymbol{x})$,

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

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* $\mathbb{E}w^2 < \infty$, *the* ***conditional variance*** *of* w *given* $\boldsymbol{x}$ *is*

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

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* $\mathbb{E}e^2 < \infty$, *the* ***conditional variance*** *of the regression error* e *is*

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

Generally, $\sigma^2(\boldsymbol{x})$ 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 $\sigma^2(\boldsymbol{x}) = \operatorname{var}(y \mid \boldsymbol{x})$ 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 $\sigma^2\left(\boldsymbol{x}\right)$ 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 $m(\boldsymbol{x}) = \mathbb{E}\left(y \mid \boldsymbol{x}\right)$ 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 2.30.

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

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

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

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

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

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 $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1 x_2$, 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. It is worth noting that this symmetry is an artificial implication of the quadratic equation (2.18), and is not a general feature of nonlinear conditional means $m(x_1, x_2)$.

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, the variable *sex* typically takes only the values *man* and *woman*. 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 wages 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. $\mathbb{E}y^2 < \infty$.
2. $\mathbb{E}\left\|\boldsymbol{x}\right\|^2 < \infty$.
3. $\boldsymbol{Q_{xx}} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ *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 $\boldsymbol{Q_{xx}} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ are linearly independent, or equivalently 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

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

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

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

The minimizer

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

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 $\boldsymbol{\beta}$:

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

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.10) is

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

Rewriting (2.22) as

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

and dividing by 2, this equation takes the form

$$\boldsymbol{Q_{xy}} = \boldsymbol{Q_{xx}}\boldsymbol{\beta} \tag{2.23}$$

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

$$\boldsymbol{\beta} = \boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{xy}}$$

or

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

It is worth taking the time to understand the notation involved in the expression (2.24). $\boldsymbol{Q}_{\boldsymbol{xx}}$ is a $k \times k$ matrix and $\boldsymbol{Q}_{\boldsymbol{xy}}$ is a $k \times 1$ column vector. Therefore, alternative expressions such as $\frac{\mathbb{E}(\boldsymbol{x}y)}{\mathbb{E}(\boldsymbol{xx}')}$ or $\mathbb{E}(\boldsymbol{x}y)(\mathbb{E}(\boldsymbol{xx}'))^{-1}$ are incoherent and incorrect. We also can now see the role of Assumption 2.18.1.3. It is equivalent to assuming that $\boldsymbol{Q}_{\boldsymbol{xx}}$ has an inverse $\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}$ 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:

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

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 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

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

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

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

as claimed.

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

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

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

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

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

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

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*.

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

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*

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

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).

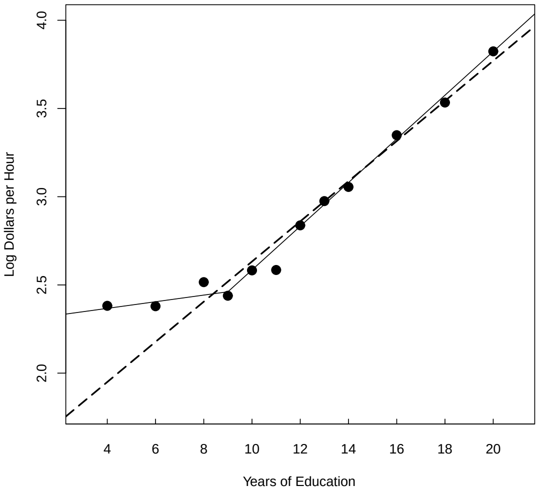


Figure 2.8: Projections of $\log(wage)$ onto Education

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

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

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:

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

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

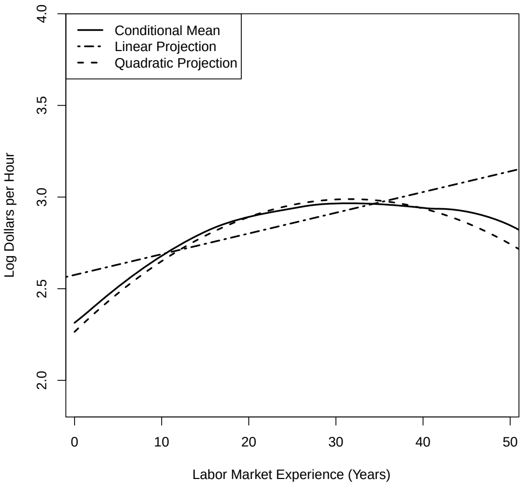


Figure 2.9: Linear and Quadratic Projections of log($wage$) onto Experience

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.

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

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

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

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

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 $\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right)$ exists and is unique as long as the $k \times k$ matrix $\boldsymbol{Q}_{\boldsymbol{xx}} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)$ 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, $\mathbb{E}\left(\boldsymbol{\alpha}'\boldsymbol{x}\right)^2 > 0$. 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 $\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right)$ 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 varibles.

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

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

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

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

where $\boldsymbol{A}^{-}$ denotes the generalized inverse of $\boldsymbol{A}$ (see Appendix A.5).

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

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

One useful feature of this formula is that it shows that $Q_{yy\cdot\boldsymbol{x}} = Q_{yy} - \boldsymbol{Q}_{y\boldsymbol{x}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{x}y}$ 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

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

or

$$\mu_y = \mu_x'\boldsymbol{\beta} + \alpha$$

where $\mu_y = \mathbb{E}y$ and $\mu_x = \mathbb{E}\boldsymbol{x}$, 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

$$\alpha = \mu_y - \mu_x'\boldsymbol{\beta}.$$

Subtracting this equation from (2.35) we find

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

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

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

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

$$\alpha = \mu_y - \mu_x'\boldsymbol{\beta} \tag{2.37}$$

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

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

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

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

$$\boldsymbol{Q}_{xx} = \left[\begin{array}{cc} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{array}\right] = \left[\begin{array}{cc} \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_1') & \mathbb{E}(\boldsymbol{x}_1\boldsymbol{x}_2') \\ \mathbb{E}(\boldsymbol{x}_2\boldsymbol{x}_1') & \mathbb{E}(\boldsymbol{x}_2\boldsymbol{x}_2') \end{array}\right]$$

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

$$\boldsymbol{Q}_{xy} = \left[\begin{array}{c} \boldsymbol{Q}_{1y} \\ \boldsymbol{Q}_{2y} \end{array}\right] = \left[\begin{array}{c} \mathbb{E}(\boldsymbol{x}_1 y) \\ \mathbb{E}(\boldsymbol{x}_2 y) \end{array}\right].$$

By the partitioned matrix inversion formula (A.4)

$$\boldsymbol{Q}_{xx}^{-1} = \left[\begin{array}{cc} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{array}\right]^{-1} \overset{def}{=} \left[\begin{array}{cc} \boldsymbol{Q}^{11} & \boldsymbol{Q}^{12} \\ \boldsymbol{Q}^{21} & \boldsymbol{Q}^{22} \end{array}\right] = \left[\begin{array}{cc} \boldsymbol{Q}_{11\cdot2}^{-1} & -\boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot1}^{-1} \end{array}\right]. \tag{2.41}$$

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

$$\begin{aligned} \boldsymbol{\beta} &= \left(\begin{array}{c} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{array}\right) \\ &= \left[\begin{array}{cc} \boldsymbol{Q}_{11\cdot2}^{-1} & -\boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot1}^{-1} \end{array}\right]\left[\begin{array}{c} \boldsymbol{Q}_{1y} \\ \boldsymbol{Q}_{2y} \end{array}\right] \\ &= \left(\begin{array}{c} \boldsymbol{Q}_{11\cdot2}^{-1}\left(\boldsymbol{Q}_{1y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2y}\right) \\ \boldsymbol{Q}_{22\cdot1}^{-1}\left(\boldsymbol{Q}_{2y} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{1y}\right) \end{array}\right) \\ &= \left(\begin{array}{c} \boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{1y\cdot2} \\ \boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{2y\cdot1} \end{array}\right) \end{aligned}$$

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 $\boldsymbol{x}_2$:

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

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

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

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

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

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

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

where

$$\boldsymbol{\Gamma} = \left(\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_1'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_1\boldsymbol{x}_2'\right)$$

is the coefficient matrix from a projection of $\boldsymbol{x}_2$ on $\boldsymbol{x}_1$.

Observe that $\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1 + \boldsymbol{\Gamma}\boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_1$ unless $\boldsymbol{\Gamma} = \boldsymbol{0}$ or $\boldsymbol{\beta}_2 = \boldsymbol{0}$. 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 $\boldsymbol{\Gamma}\boldsymbol{\beta}_2$ 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 $\Gamma > 0$. 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 $\Gamma\beta_2 > 0$ and $\gamma_1 = \beta_1 + \Gamma\beta_2 > \beta_1$. 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.

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 $m(\boldsymbol{x})$

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

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

$$\begin{aligned} \boldsymbol{\beta} &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x}m(\boldsymbol{x})\right) && (2.46) \\ &= \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1} \mathbb{E}\left(\boldsymbol{x}y\right) && (2.47) \end{aligned}$$

(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 Normal Regression

Suppose the variables $(y, \boldsymbol{x})$ are jointly normally distributed. Consider the best linear predictor of y given $\boldsymbol{x}$

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

Since the error e is a linear transformation of the normal vector $(y, \boldsymbol{x})$, it follows that $(e, \boldsymbol{x})$ is jointly normal, and since they are jointly normal and uncorrelated (since $\mathbb{E}\left(\boldsymbol{x}e\right) = 0$) they are also independent (see Appendix B.9). Independence implies that

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

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

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

where

$$e \sim \mathrm{N}(0, \sigma^2)$$

is independent of $\boldsymbol{x}$.

This is an alternative (and traditional) motivation for the linear CEF model. This motivation has limited merit in econometric applications since economic data is typically non-normal.

2.26 Regression to the Mean

The term **regression** originated in an influential paper by Francis Galton published in 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 (B.7) 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.

2.27 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.

2.28 Limitations of the Best Linear Predictor

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 a 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 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 – 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 apparant difference is a by-product of a linear approximation to a non-linear mean, combined with different marginal distributions for the conditioning variables.

2.29 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 $m(x) = 2x - x^2/6$.

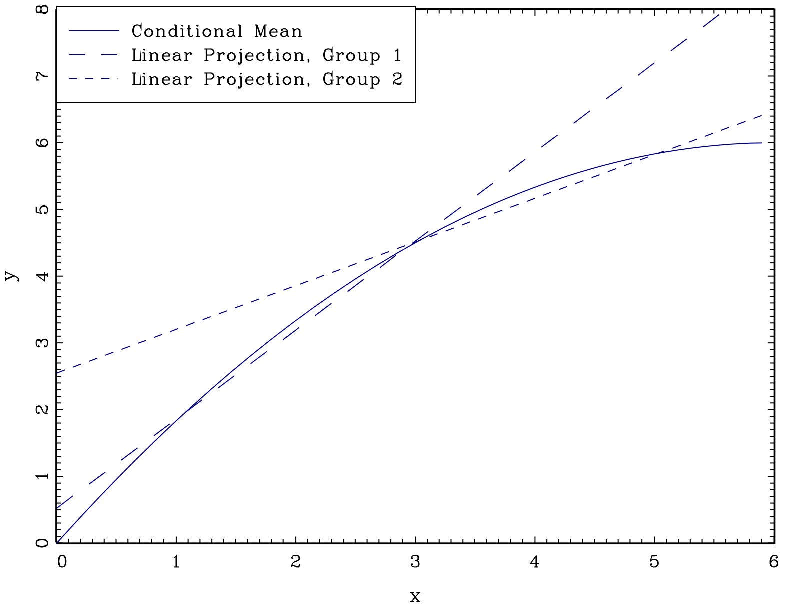


Figure 2.10: Conditional Mean and Two Linear Projections

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 $\boldsymbol{\eta} = \boldsymbol{\nabla}\left(\boldsymbol{x}'\boldsymbol{\eta}\right)$ 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 $\boldsymbol{\eta}$ **:**

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

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 2.29.1 *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*

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

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

2.30 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 includes as a special case the random coefficient model

(2.29) 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 2.30.1 *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 2.30.2 *In the model (2.52) the* ***average causal effect*** *of x_1 on y conditional on $\boldsymbol{x}_2$ is*

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

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

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

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 2.30.3 ***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 2.30.1 *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 attendence takes the form

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

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

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

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

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 attendence, 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 2.30.2 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 attendence 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 cofficient on *college*, \$5.50, is the regression derivative of college attendence for those with low test scores, and the sum of this coefficient with the interaction coefficient, \$8.50, is the regression derivative for college attendence for those with high test scores. These equal the average causal effect.

In this example, by conditioning on the aptitude test score, the average causal effect of education on wages can be learned from a regression analyis. 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.

2.31 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

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

and if y is continuous with density f then

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

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

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

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

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

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

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

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

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

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

By Liapunov's Inequality (B.20), (2.59) implies $\mathbb{E}\left|y\right|^s < \infty$ 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.

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

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

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 show the Pareto density for $a = 1$ and $a = 2$. The parameter a also determines which moments are finite. We can calculate that

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

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.

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 $r < a$

$$\mathbb{E}\left|y\right|^r = \int_{-\infty}^{\infty} \left|y\right|^r f(y)dy \leq \int_{-1}^{1} f(y)dy + 2A\int_1^{\infty} y^{r-a-1}dy \leq 1 + \frac{2A}{a-r} < \infty.$$

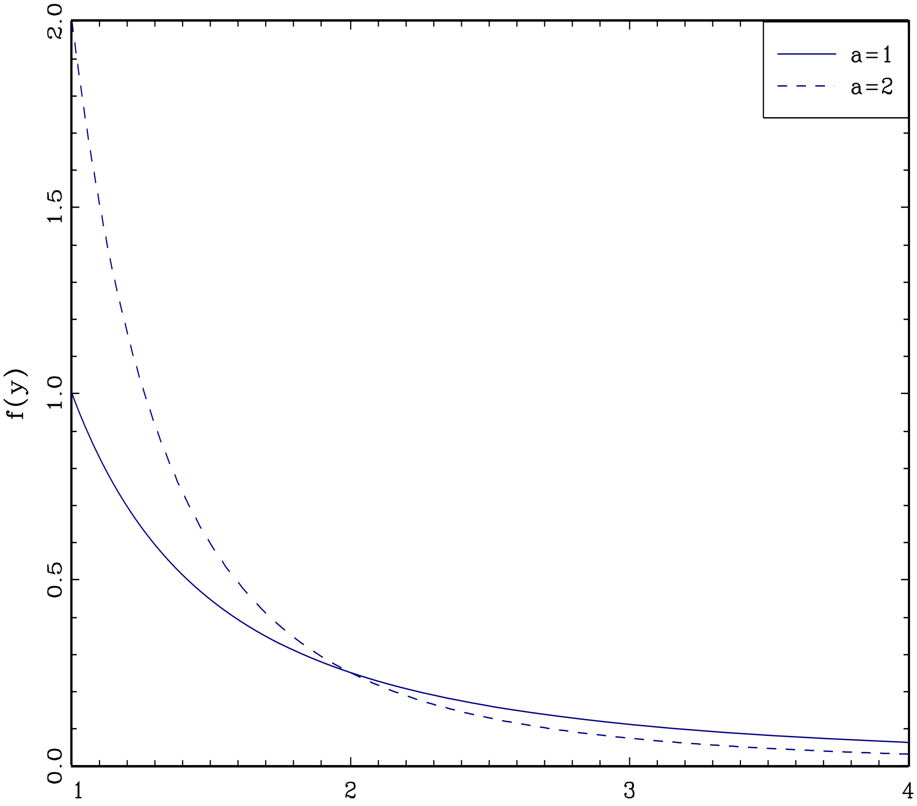


Figure 2.11: Pareto Densities, $a = 1$ and $a = 2$

Thus if the tail of the density declines at the rate $|y|^{-a-1}$ 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 2.31.1 *For any non-negative random variable y*

$$\mathbb{E}y = \int_0^\infty \Pr(y > u)\, du$$

Proof of Theorem 2.31.1: Let $F^*(x) = \Pr(y > x) = 1 - F(x)$, where $F(x)$ is the distribution function. By integration by parts

$$\mathbb{E}y = \int_0^\infty ydF(y) = -\int_0^\infty ydF^*(y) = -\left[yF^*(y)\right]_0^\infty + \int_0^\infty F^*(y)dy = \int_0^\infty \Pr(y > u)\, du$$

as stated. ■

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 measurable sets $\mathcal{X}$

$$\mathbb{E}\left(1\left(\boldsymbol{x} \in \mathcal{X}\right) y\right) = \mathbb{E}\left(1\left(\boldsymbol{x} \in \mathcal{X}\right) m(\boldsymbol{x})\right). \tag{2.60}$$

The function $m(\boldsymbol{x})$ is almost everywhere unique, in the sense that if $h(\boldsymbol{x})$ satisfies (2.60), 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 $m(\boldsymbol{x}) = \mathbb{E}\left(y \mid \boldsymbol{x}\right)$.*

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

The conditional mean $m(\boldsymbol{x})$ defined by (2.60) specializes to (2.7) when $(y, \boldsymbol{x})$ have a joint density. The usefulness of definition (2.60) 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.

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 equalling 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

$$f(y|\boldsymbol{x}_1, \boldsymbol{x}_2) f(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{f(y, \boldsymbol{x}_1, \boldsymbol{x}_2)}{f(\boldsymbol{x}_1, \boldsymbol{x}_2)} \frac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f(\boldsymbol{x}_1)} = f(y, \boldsymbol{x}_2|\boldsymbol{x}_1), \tag{2.61}$$

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

$$\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \int_{\mathbb{R}} y f(y|\boldsymbol{x}_1, \boldsymbol{x}_2)\, dy. \tag{2.62}$$

Integrating (2.62) with respect to the conditional density of $\boldsymbol{x}_2$ given $\boldsymbol{x}_1$, and applying (2.61) 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 (2.9). The assumption that $\mathbb{E}|g(\boldsymbol{x}) y| < \infty$ is required for the first equality to be well-defined. Equation (2.10) follows by applying the Simple Law of Iterated Expectations to (2.9). ■

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

Set $z = \mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)$. By the conditional Jensen's inequality (B.13),

$$(\mathbb{E}(z \mid \boldsymbol{x}_1))^2 \le \mathbb{E}\left(z^2 \mid \boldsymbol{x}_1\right).$$

Taking unconditional expectations, this implies

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

Similarly,

$$(\mathbb{E}y)^2 \le \mathbb{E}\left((\mathbb{E}(y \mid \boldsymbol{x}_1))^2 \right) \le \mathbb{E}\left((\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2))^2 \right). \tag{2.63}$$

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 (2.63) implies that the variances are ranked monotonically:

$$0 \leq \operatorname{var}\left(\mathbb{E}(y \mid \boldsymbol{x}_1)\right) \leq \operatorname{var}\left(\mathbb{E}(y \mid \boldsymbol{x}_1, \boldsymbol{x}_2)\right). \tag{2.64}$$

Next, for $\mu = \mathbb{E}y$ observe that

$$\mathbb{E}\left(y - \mathbb{E}(y \mid \boldsymbol{x})\right)\left(\mathbb{E}(y \mid \boldsymbol{x}) - \mu\right) = \mathbb{E}\left(y - \mathbb{E}(y \mid \boldsymbol{x})\right)\left(\mathbb{E}(y \mid \boldsymbol{x}) - \mu\right) = 0$$

so the decomposition

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

satisfies

$$\operatorname{var}(y) = \operatorname{var}\left(y - \mathbb{E}(y \mid \boldsymbol{x})\right) + \operatorname{var}\left(\mathbb{E}(y \mid \boldsymbol{x})\right). \tag{2.65}$$

The monotonicity of the variances of the conditional mean (2.64) applied to the variance decomposition (2.65) implies the reverse monotonicity of the variances of the differences, completing the proof. ■

Proof of Theorem 2.8.1. Applying Minkowski's Inequality (B.19) to $e = y - m(\boldsymbol{x})$,

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

where the two parts on the right-hand are finite since $\mathbb{E}\left|y\right|^r < \infty$ by assumption and $\mathbb{E}\left|m(\boldsymbol{x})\right|^r < \infty$ by the Conditional Expectation Inequality (B.14). The fact that $\left(\mathbb{E}\left|e\right|^r\right)^{1/r} < \infty$ implies $\mathbb{E}\left|e\right|^r < \infty$. ■

Proof of Theorem 2.18.1. For part 1, by the Expectation Inequality (B.15), (A.19) and Assumption 2.18.1,

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

Similarly, using the Expectation Inequality (B.15), the Cauchy-Schwarz Inequality (B.17) and Assumption 2.18.1,

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

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

For part 2, the coefficient $\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right)$ 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

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

The first inequality holds because $\mathbb{E}\left(y\boldsymbol{x}'\right)\left(\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}y\right)$ is a quadratic form and therefore necessarily non-negative. Second, by the Expectation Inequality (B.15), the Cauchy-Schwarz Inequality (B.17) and Assumption 2.18.1,

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

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

For part 6, Applying Minkowski's Inequality (B.19) 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 $\mathbb{E}(y \mid x) = a + bx$, find $\mathbb{E}(yx)$ 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 $\mathbb{E}(y \mid x)$, $\mathbb{E}\left(y^2 \mid x\right)$ and $\operatorname{var}(y \mid x)$ 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

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

Exercise 2.7 Show that the conditional variance can be written as

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

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:

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

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

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

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 $\mathbb{E}(y \mid x_1, x_2)$ as a linear regression.

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

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

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

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 $0 \le x \le 1$, $0 \le y \le 1$. 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 $\sigma^2 = \operatorname{var}(x)$. 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

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

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

Exercise 2.20 Verify that (2.60) holds with $m(\boldsymbol{x})$ defined in (2.7) when $(y, \boldsymbol{x})$ have a joint density $f(y, \boldsymbol{x})$.

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 defered to later chapters.

3.2 Random 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

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

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 emphasize when we are discussing observations. Typically in econometrics we 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 $\{(y_i, \boldsymbol{x}_i); i = 1, ..., n\}$.

From the viewpoint of empirical analysis, a dataset is a 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. For cross-sectional studies, the most common approach is to treat the individual observations as independent draws from an underlying population F. When the observations are realizations of independent and identically distributed random variables, we say that the data is a random sample.

Assumption 3.2.1 *The observations* $\{(y_1, \boldsymbol{x}_1), ..., (y_i, \boldsymbol{x}_i), ..., (y_n, \boldsymbol{x}_n)\}$ *are a random sample.*

With a random sample, the ordering of the data is irrelevant. There is nothing special about any specific observation or ordering. You can permute the order of the observations and no information is gained or lost.

As most economic data sets are not literally the result of a random experiment, the random sampling framework is best viewed as an approximation rather than being literally true.

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

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{3.1}$$

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

$$\boldsymbol{\beta} = \underset{b \in \mathbb{R}^k}{\operatorname{argmin}}\, S(\boldsymbol{b}), \tag{3.2}$$

the minimizer of the expected squared error

$$S(\boldsymbol{\beta}) = \mathbb{E}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2, \tag{3.3}$$

and has the explicit solution

$$\boldsymbol{\beta} = \left(\mathbb{E}\left(\boldsymbol{x_i}\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{x}_i y_i\right). \tag{3.4}$$

3.3 Sample Means

Consider the intercept-only model

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

In this case the regression parameter is the unconditional mean $\mu = \mathbb{E}(y_i)$.

The standard estimator of a population mean is the sample mean, namely

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

The sample mean is the empirical analog of the population mean, and is the conventional estimator in the lack of other information about μ or the distribution of y. We call $\widehat{\mu}$ the **moment estimator** for μ.

Indeed, whenever we have a parameter which can be written as the expectation of a function of random variables, a natural estimator of the parameter is the moment estimator, which is the sample mean of the corresponding function of the observations. For example, for $\mu_2 = \mathbb{E}(y_i^2)$ the moment estimator is $\widehat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n y_i^2$, and for $\theta = \mathbb{E}(y_{1i}y_{2i})$ the moment estimator is $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^n y_{1i}y_{2i}$.

3.4 Least Squares Estimator

The linear projection coefficient $\boldsymbol{\beta}$ is defined in (3.2) as the minimizer of the expected squared error $S(\boldsymbol{\beta})$ defined in (3.3). 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:

$$\begin{aligned} S_n(\boldsymbol{\beta}) &= \frac{1}{n}\sum_{i=1}^n \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \qquad (3.5) \\ &= \frac{1}{n} SSE_n(\boldsymbol{\beta}) \end{aligned}$$

where

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

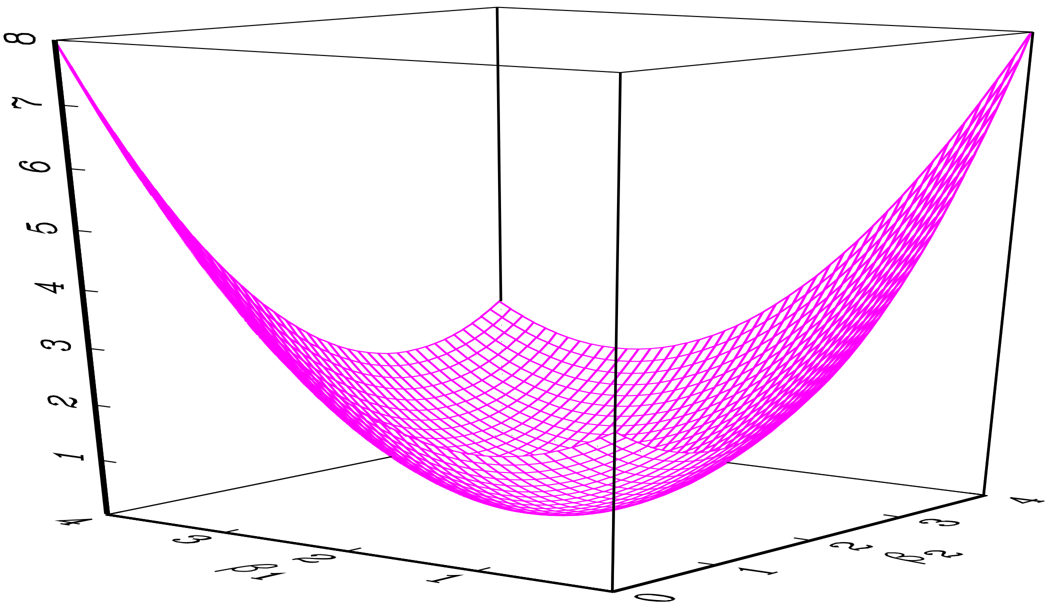


Figure 3.1: Sum-of-Squared Errors Function

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

Since $S_n(\boldsymbol{\beta})$ is a sample average, we can interpret it as an estimator of the expected squared error $S(\boldsymbol{\beta})$. Examining $S_n(\boldsymbol{\beta})$ as a function of $\boldsymbol{\beta}$ therefore is informative about how $S(\boldsymbol{\beta})$ varies with $\boldsymbol{\beta}$. The projection coefficient coeffient minimizes $S(\boldsymbol{\beta})$, an analog estimator minimizes (3.5):

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

Alternatively, as $S_n(\boldsymbol{\beta})$ is a scale multiple of $SSE_n(\boldsymbol{\beta})$, 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) (or ordinary least squares (OLS)) estimator** of $\boldsymbol{\beta}$. 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, as just by seeing the symbol $\widehat{\boldsymbol{\beta}}$ we can immediately interpret it as an estimator (because of the hat), and as an estimator of a parameter labelled $\boldsymbol{\beta}$. 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 $S_n(\boldsymbol{\beta})$, Figure 3.1 displays an example sum-of-squared errors function $SSE_n(\boldsymbol{\beta})$ for the case $k = 2$. The least-squares estimator $\widehat{\boldsymbol{\beta}}$ is the the pair $(\widehat{\beta}_1, \widehat{\beta}_2)$ minimizing 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

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

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 $SSE_n(\beta)$ is

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}. \tag{3.6}$$

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

$$\widehat{\beta} = \frac{\sum_{i=1}^n 1y_i}{\sum_{i=1}^n 1^2} = \frac{1}{n}\sum_{i=1}^n y_i = \overline{y}, \tag{3.7}$$

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

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

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

$$0 = \frac{\partial}{\partial\boldsymbol{\beta}}SSE_n(\widehat{\boldsymbol{\beta}}) = -2\sum_{i=1}^n \boldsymbol{x}_i y_i + 2\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}. \tag{3.8}$$

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 (3.8). 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

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_i y_i\right). \tag{3.9}$$

This is the natural estimator of the best linear projection coefficient $\boldsymbol{\beta}$ defined in (3.2), and can also be called the linear projection estimator.

We see that (3.9) simplifies to the expression (3.6) when $k = 1$. The expression (3.9) is a notationally simple generalization but requires a careful attention to vector and matrix manipulations.

Alternatively, equation (3.4) writes the projection coefficient $\boldsymbol{\beta}$ as an explicit function of the population moments $\boldsymbol{Q}_{xy}$ and $\boldsymbol{Q}_{xx}$. Their moment estimators are the sample moments

$$\begin{aligned}\widehat{\boldsymbol{Q}}_{xy} &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i y_i \\ \widehat{\boldsymbol{Q}}_{xx} &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'.\end{aligned}$$

The moment estimator of $\boldsymbol{\beta}$ replaces the population moments in (3.4) with the sample moments:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{Q}}_{xx}^{-1}\widehat{\boldsymbol{Q}}_{xy} \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right) \\
&= \left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i y_i\right)
\end{aligned}
$$

which is identical with (3.9).

Least Squares Estimation

Definition 3.6.1 *The* ***least-squares estimator*** $\widehat{\boldsymbol{\beta}}$ *is*

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

where

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

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, (3.1) 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 (3.8) 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 generate the estimates from 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 wages 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

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix} \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix} \\ &= \begin{pmatrix} 0.155 \\ 0.698 \end{pmatrix}. \end{aligned} \tag{3.10}$$

We often write the estimated equation using the format

$$\widehat{\log(Wage)} = 0.155 \; education + 0.698. \tag{3.11}$$

An interpretation of the estimated equation is that each year of education is associated with an 16% increase in mean wages.

Equation (3.11) 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 (3.11) 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

$$\widehat{\log(Wage)} = 0.143 \; education + 0.036 \; experience - 0.071 \; experience^2/100 + 0.575. \tag{3.12}$$

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**

$$\hat{y}_i = \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}$$

and the **residual**

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}. \tag{3.13}$$

Sometimes $\hat{y}_i$ is called the predicted value, but this is a misleading label. The fitted value $\hat{y}_i$ 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 $\hat{y}_i$ as a *fitted* rather than a *predicted* value.

Note that $y_i = \hat{y}_i + \hat{e}_i$ and

$$y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \hat{e}_i. \tag{3.14}$$

We make a distinction between the **error** e_i and the **residual** $\hat{e}_i$. The error e_i is unobservable while the residual $\hat{e}_i$ is a by-product of estimation. These two variables are frequently mislabeled, which can cause confusion.

Equation (3.8) implies that

$$\sum_{i=1}^{n} \boldsymbol{x}_i\hat{e}_i = \mathbf{0}. \tag{3.15}$$

To see this by a direct calculation, using (3.13) and (3.9),

$$\begin{aligned}
\sum_{i=1}^{n} \boldsymbol{x}_i\hat{e}_i &= \sum_{i=1}^{n} \boldsymbol{x}_i\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right) \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i y_i\right) \\
&= \sum_{i=1}^{n} \boldsymbol{x}_i y_i - \sum_{i=1}^{n} \boldsymbol{x}_i y_i \\
&= \mathbf{0}.
\end{aligned}$$

When $\boldsymbol{x}_i$ contains a constant, an implication of (3.15) is

$$\frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0. \tag{3.16}$$

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 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

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{3.17}$$

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

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}\right). \tag{3.18}$$

The matrix version of (3.14) and estimated version of (3.17) is

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}},$$

or equivalently the residual vector is

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}.$$

Using the residual vector, we can write (3.15) as

$$\boldsymbol{X}'\hat{\boldsymbol{e}} = \boldsymbol{0}. \tag{3.20}$$

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

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{y}\right) \\ \hat{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ \boldsymbol{X}'\hat{\boldsymbol{e}} &= \boldsymbol{0}. \end{aligned}$$

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.

3.10 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** and **idempotent**[1]. 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:

$$\boldsymbol{P}\boldsymbol{y}=\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}=\boldsymbol{X}\widehat{\boldsymbol{\beta}}=\hat{\boldsymbol{y}}.$$

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

$$\begin{aligned}\boldsymbol{P}_1\boldsymbol{y} &= \boldsymbol{1}\left(\boldsymbol{1}'\boldsymbol{1}\right)^{-1}\boldsymbol{1}'\boldsymbol{y} \\ &= \boldsymbol{1}\bar{y}\end{aligned}$$

creates an n-vector whose elements are the sample mean $\bar{y}$ of y_i.

[1] A matrix $\boldsymbol{P}$ is symmetric if $\boldsymbol{P}'=\boldsymbol{P}$. A matrix $\boldsymbol{P}$ is idempotent if $\boldsymbol{P}\boldsymbol{P}=\boldsymbol{P}$. See Appendix A.8.

The i'th diagonal element of $\boldsymbol{P}=\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ is

$$h_{ii}=\boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i \tag{3.21}$$

which is called the **leverage** of the i'th observation.

Some useful properties of the the matrix $\boldsymbol{P}$ and the leverage values h_{ii} are now summarized.

Theorem 3.10.1

$$\sum_{i=1}^{n} h_{ii}=\operatorname{tr}\boldsymbol{P}=k \tag{3.22}$$

and

$$0\leq h_{ii}\leq 1. \tag{3.23}$$

To show (3.22),

$$\begin{aligned}\operatorname{tr}\boldsymbol{P}&=\operatorname{tr}\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right)\\&=\operatorname{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\right)\\&=\operatorname{tr}\left(\boldsymbol{I}_k\right)\\&=k.\end{aligned}$$

See Appendix A.4 for definition and properties of the trace operator. The proof of (3.23) is defered to Section 3.21.

3.11 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

$$\boldsymbol{M}\boldsymbol{X}=\left(\boldsymbol{I}_n-\boldsymbol{P}\right)\boldsymbol{X}=\boldsymbol{X}-\boldsymbol{P}\boldsymbol{X}=\boldsymbol{X}-\boldsymbol{X}=\boldsymbol{0}.$$

Thus $\boldsymbol{M}$ and $\boldsymbol{X}$ are orthogonal. We call $\boldsymbol{M}$ an **orthogonal projection matrix** or 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 many 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 (3.22) we can calculate

$$\operatorname{tr}\boldsymbol{M}=n-k. \tag{3.24}$$

(See Exercise 3.9.) While $\boldsymbol{P}$ creates fitted values, $\boldsymbol{M}$ creates least-squares residuals:

$$\boldsymbol{M}\boldsymbol{y}=\boldsymbol{y}-\boldsymbol{P}\boldsymbol{y}=\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}=\widehat{\boldsymbol{e}}. \tag{3.25}$$

As discussed in the previous section, a special example of a projection matrix occurs when $\boldsymbol{X} = \mathbf{1}$ is an n-vector of ones, so that $\boldsymbol{P}_1 = \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'$. Similarly, set

$$\begin{aligned}\boldsymbol{M}_1 &= \boldsymbol{I}_n - \boldsymbol{P}_1 \\ &= \boldsymbol{I}_n - \mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)^{-1}\mathbf{1}'.\end{aligned}$$

While $\boldsymbol{P}_1$ creates a vector of sample means, $\boldsymbol{M}_1$ creates demeaned values:

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{y} - \mathbf{1}\bar{y}.$$

For simplicity we will often write the right-hand-side as $\boldsymbol{y} - \bar{y}$. The i'th element is $y_i - \bar{y}$, the **demeaned** value of y_i.

We can also use (3.25) to write an alternative expression for the residual vector. Substituting $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ into $\hat{\boldsymbol{e}} = \boldsymbol{My}$ and using $\boldsymbol{MX} = \mathbf{0}$ we find

$$\hat{\boldsymbol{e}} = \boldsymbol{My} = \boldsymbol{M}\left(\boldsymbol{X\beta} + \boldsymbol{e}\right) = \boldsymbol{Me} \tag{3.26}$$

which is free of dependence on the regression coefficient $\boldsymbol{\beta}$.

3.12 Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}e_i^2$ is a moment, so a natural estimator is a moment estimator. If e_i were observed we would estimate σ^2 by

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{3.27}$$

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 $\hat{e}_i$ are calculated in the first step, and then we substitute $\hat{e}_i$ for e_i in expression (3.27) to obtain the feasible estimator

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2. \tag{3.28}$$

In matrix notation, we can write (3.27) and (3.28) as

$$\tilde{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e}$$

and

$$\hat{\sigma}^2 = n^{-1}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}. \tag{3.29}$$

Recall the expressions $\widehat{\boldsymbol{e}} = \boldsymbol{My} = \boldsymbol{Me}$ from (3.25) and (3.26). Applied to (3.29) we find

$$\begin{aligned}\hat{\sigma}^2 &= n^{-1}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \\ &= n^{-1}\boldsymbol{y}'\boldsymbol{MMy} \\ &= n^{-1}\boldsymbol{y}'\boldsymbol{My} \\ &= n^{-1}\boldsymbol{e}'\boldsymbol{Me}\end{aligned}$$

the third equality since $\boldsymbol{MM} = \boldsymbol{M}$.

An interesting implication is that

$$\begin{aligned}\tilde{\sigma}^2 - \hat{\sigma}^2 &= n^{-1}\boldsymbol{e}'\boldsymbol{e} - n^{-1}\boldsymbol{e}'\boldsymbol{Me} \\ &= n^{-1}\boldsymbol{e}'\boldsymbol{Pe} \\ &\geq 0.\end{aligned}$$

The final inequality holds because $\boldsymbol{P}$ is positive semi-definite and $\boldsymbol{e}'\boldsymbol{Pe}$ is a quadratic form. This shows that the feasible estimator $\hat{\sigma}^2$ is numerically smaller than the idealized estimator (3.27).

3.13 Analysis of Variance

Another way of writing (3.25) is

$$\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} + \boldsymbol{M}\boldsymbol{y} = \hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}. \tag{3.30}$$

This decomposition is **orthogonal**, that is

$$\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} = (\boldsymbol{P}\boldsymbol{y})'(\boldsymbol{M}\boldsymbol{y}) = \boldsymbol{y}'\boldsymbol{P}\boldsymbol{M}\boldsymbol{y} = 0.$$

It follows that

$$\boldsymbol{y}'\boldsymbol{y} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + 2\hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{y}} + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

or

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} \hat{e}_i^2.$$

Subtracting $\bar{y}$ from both sizes of (3.30) we obtain

$$\boldsymbol{y} - \boldsymbol{1}\bar{y} = \hat{\boldsymbol{y}} - \boldsymbol{1}\bar{y} + \hat{\boldsymbol{e}}$$

This decomposition is also orthogonal when $\boldsymbol{X}$ contains a constant, as

$$(\hat{\boldsymbol{y}} - \boldsymbol{1}\bar{y})'\hat{\boldsymbol{e}} = \hat{\boldsymbol{y}}'\hat{\boldsymbol{e}} - \bar{y}\boldsymbol{1}'\hat{\boldsymbol{e}} = 0$$

under (3.16). It follows that

$$(\boldsymbol{y} - \boldsymbol{1}\bar{y})'(\boldsymbol{y} - \boldsymbol{1}\bar{y}) = (\hat{\boldsymbol{y}} - \boldsymbol{1}\bar{y})'(\hat{\boldsymbol{y}} - \boldsymbol{1}\bar{y}) + \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$$

or

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{e}_i^2.$$

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**:

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} \hat{e}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

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.

3.14 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

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}. \tag{3.31}$$

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

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}} = \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\widehat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}. \tag{3.32}$$

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 $\widehat{\boldsymbol{Q}}_{xx}$ and $\widehat{\boldsymbol{Q}}_{xy}$ as

$$\widehat{\boldsymbol{Q}}_{xx} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2 \\ \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1 & \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2 \end{bmatrix}$$

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

$$\widehat{\boldsymbol{Q}}_{xy} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1y} \\ \widehat{\boldsymbol{Q}}_{2y} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{y} \\ \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{y} \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.4)

$$\widehat{\boldsymbol{Q}}_{xx}^{-1} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix}^{-1} \overset{def}{=} \begin{bmatrix} \widehat{\boldsymbol{Q}}^{11} & \widehat{\boldsymbol{Q}}^{12} \\ \widehat{\boldsymbol{Q}}^{21} & \widehat{\boldsymbol{Q}}^{22} \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1} \end{bmatrix} \tag{3.33}$$

where $\widehat{\boldsymbol{Q}}_{11\cdot 2} = \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}$ and $\widehat{\boldsymbol{Q}}_{22\cdot 1} = \widehat{\boldsymbol{Q}}_{22} - \widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{12}$.

Thus

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} \\ &= \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{Q}}_{1y} \\ \widehat{\boldsymbol{Q}}_{2y} \end{bmatrix} \\ &= \begin{pmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{1y\cdot 2} \\ \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{2y\cdot 1} \end{pmatrix} \end{aligned}$$

Now

$$\begin{aligned} \widehat{\boldsymbol{Q}}_{11\cdot 2} &= \widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21} \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_1 - \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{X}_2\left(\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\frac{1}{n}\boldsymbol{X}_2'\boldsymbol{X}_1 \\ &= \frac{1}{n}\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1 \end{aligned}$$

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 $\widehat{\boldsymbol{Q}}_{22\cdot 1} = \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2$ 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 $\widehat{\boldsymbol{Q}}_{2y\cdot 1} = \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}$.

Therefore

$$\widehat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{y}\right) \tag{3.34}$$

and

$$\widehat{\boldsymbol{\beta}}_2 = \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}\right). \tag{3.35}$$

These are algebraic expressions for the sub-coefficient estimates from (3.32).

3.15 Residual Regression

As first recognized by Frisch and Waugh (1933), expressions (3.34) and (3.35) 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 (3.35). Since $\boldsymbol{M}_1$ is idempotent, $\boldsymbol{M}_1 = \boldsymbol{M}_1\boldsymbol{M}_1$ and thus

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_2 &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{y}\right) \\ &= \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{M}_1\boldsymbol{y}\right) \\ &= \left(\widetilde{\boldsymbol{X}}_2'\widetilde{\boldsymbol{X}}_2\right)^{-1}\left(\widetilde{\boldsymbol{X}}_2'\tilde{\boldsymbol{e}}_1\right)\end{aligned}$$

where

$$\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$$

and

$$\tilde{\boldsymbol{e}}_1 = \boldsymbol{M}_1\boldsymbol{y}.$$

Thus the coefficient estimate $\widehat{\boldsymbol{\beta}}_2$ is algebraically equal to the least-squares regression of $\tilde{\boldsymbol{e}}_1$ 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 $\tilde{\boldsymbol{e}}_1$ 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 3.15.1 ***Frisch-Waugh-Lovell***
In the model (3.31), the OLS estimator of $\boldsymbol{\beta}_2$ and the OLS residuals $\hat{\boldsymbol{e}}$ may be equivalently computed by either the OLS regression (3.32) or via the following algorithm:

1. *Regress $\boldsymbol{y}$ on $\boldsymbol{X}_1$, obtain residuals $\tilde{\boldsymbol{e}}_1$;*

2. *Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, obtain residuals $\widetilde{\boldsymbol{X}}_2$;*

3. *Regress $\tilde{\boldsymbol{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$, obtain OLS estimates $\widehat{\boldsymbol{\beta}}_2$ and residuals $\hat{\boldsymbol{e}}$.*

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 (3.32), 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 intrepreting regression coefficients.

A common application of the FWL theorem, which you may have seen in an introductory econometrics course, is the demeaning formula for regression. 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 $\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2$:

$$\widehat{\boldsymbol{\beta}}_2 = \left(\sum_{i=1}^{n}\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)'\right)^{-1}\left(\sum_{i=1}^{n}\left(\boldsymbol{x}_{2i} - \overline{\boldsymbol{x}}_2\right)\left(y_i - \overline{y}\right)\right).$$

Thus the OLS estimator for the slope coefficients is a regression with demeaned data.

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.

3.16 Prediction Errors

The least-squares residual $\hat{e}_i$ 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:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(-i)} &= \left(\frac{1}{n-1}\sum_{j\neq i}\boldsymbol{x}_j\boldsymbol{x}_j'\right)^{-1}\left(\frac{1}{n-1}\sum_{j\neq i}\boldsymbol{x}_j y_j\right)\\ &= \left(\boldsymbol{X}_{(-i)}'\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}_{(-i)}\boldsymbol{y}_{(-i)}. \qquad (3.36)\end{aligned}$$

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

$$\tilde{y}_i = \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_{(-i)},$$

and the **leave-one-out residual** or **prediction error** or **prediction residual** is

$$\tilde{e}_i = y_i - \tilde{y}_i.$$

A convenient alternative expression for $\widehat{\boldsymbol{\beta}}_{(-i)}$ (derived in Section 3.21) is

$$\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}} - (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \tag{3.37}$$

where h_{ii} are the leverage values as defined in (3.21).

Using (3.37) we can simplify the expression for the prediction error:

$$\begin{aligned} \tilde{e}_i &= y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_{(-i)} \\ &= y_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}} + (1 - h_{ii})^{-1} \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \\ &= \hat{e}_i + (1 - h_{ii})^{-1} h_{ii} \hat{e}_i \\ &= (1 - h_{ii})^{-1} \hat{e}_i. \end{aligned} \tag{3.38}$$

To write this in vector notation, define

$$\begin{aligned} \boldsymbol{M}^* &= \left(\boldsymbol{I}_n - \operatorname{diag}\{h_{11}, .., h_{nn}\}\right)^{-1} \\ &= \operatorname{diag}\{(1 - h_{11})^{-1}, .., (1 - h_{nn})^{-1}\}. \end{aligned} \tag{3.39}$$

Then (3.38) is equivalent to

$$\widetilde{\boldsymbol{e}} = \boldsymbol{M}^* \widehat{\boldsymbol{e}}. \tag{3.40}$$

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

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{n} \sum_{i=1}^{n} \tilde{e}_i^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} (1 - h_{ii})^{-2} \hat{e}_i^2. \end{aligned} \tag{3.41}$$

This is also known as the **sample mean squared prediction error**. Its square root $\tilde{\sigma} = \sqrt{\tilde{\sigma}^2}$ is the **prediction standard error**.

3.17 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

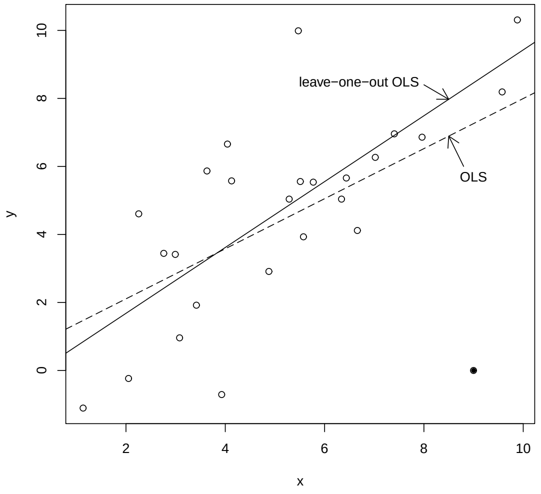


Figure 3.2: Impact of an influential observation on the least-squares estimator

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 (3.37)-(3.38) we know that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{\beta}}_{(-i)} &= (1-h_{ii})^{-1}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\hat{e}_i \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\tilde{e}_i. \end{aligned} \tag{3.42}$$

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

$$\begin{aligned}\hat{y}_i-\tilde{y}_i &= \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} \\ &= \boldsymbol{x}_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\tilde{e}_i \\ &= h_{ii}\tilde{e}_i\end{aligned}$$

which is a simple function of the leverage values h_{ii} and prediction errors $\tilde{e}_i$. Observation i is influential for the predicted value if $|h_{ii}\tilde{e}_i|$ is large, which requires that both h_{ii} and $|\tilde{e}_i|$ 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 $\tilde{e}_i$ 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 (3.42). 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

$$Influence = \max_{1\leq i\leq n} |\hat{y}_i - \tilde{y}_i| = \max_{1\leq i\leq n} |h_{ii}\tilde{e}_i| .$$

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 an fair amount of individual judgement. 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 (3.12) 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 twelth 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 to his unusual characteristics (very low education and very high experience) within this sample. Essentially, regression (3.12) 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.

$$\widehat{\log(Wage)} = 0.144\ education + 0.043\ experience - 0.095\ experience^2/100 + 0.531. \tag{3.43}$$

For this regression, we calculate that $Influence = 0.11$, which is greatly reduced relative to the regression (3.12). Comparing (3.43) with (3.12), 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 (3.43) can be viewed as a more robust estimate of the conditional mean for most levels of *experience*. Whether to report (3.12) or (3.43) in an application is largely a matter of judgment.

3.18 Normal Regression Model

The normal regression model is the linear regression model under the restriction that the error e_i is independent of $\boldsymbol{x}_i$ and has the distribution $\mathrm{N}\left(0,\sigma^2\right)$. We can write this as

$$e_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(0,\sigma^2\right).$$

This assumption implies

$$y_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2\right).$$

Normal regression is a parametric model, where likelihood methods can be used for estimation, testing, and distribution theory.

The log-likelihood function for the normal regression model is

$$\begin{aligned}\log L(\boldsymbol{\beta}, \sigma^2) &= \sum_{i=1}^{n} \log\left(\frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2\right)\right) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2} SSE_n(\boldsymbol{\beta}). \qquad (3.44)\end{aligned}$$

The maximum likelihood estimator (MLE) $(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}^2_{\text{mle}})$ maximizes $\log L(\boldsymbol{\beta}, \sigma^2)$. Since the latter is a function of $\boldsymbol{\beta}$ only through the sum of squared errors $SSE_n(\boldsymbol{\beta})$, maximizing the likelihood is identical to minimizing $SSE_n(\boldsymbol{\beta})$. Hence

$$\widehat{\boldsymbol{\beta}}_{\text{mle}} = \widehat{\boldsymbol{\beta}}_{\text{ols}},$$

the MLE for $\boldsymbol{\beta}$ equals the OLS estimator. Due to this equivalence, the least squares estimator $\widehat{\boldsymbol{\beta}}_{\text{ols}}$ is often called the MLE.

We can also find the MLE for σ^2. Plugging $\widehat{\boldsymbol{\beta}}$ into the log-likelihood we obtain

$$\log L\left(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \sigma^2\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{SSE_n(\widehat{\boldsymbol{\beta}}_{\text{mle}})}{2\sigma^2}.$$

Maximization with respect to σ^2 yields the first-order condition

$$\frac{\partial}{\partial\sigma^2}\log L\left(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}^2\right) = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\left(\hat{\sigma}^2\right)^2} SSE_n(\widehat{\boldsymbol{\beta}}_{\text{mle}}) = 0.$$

Solving for $\hat{\sigma}^2$ yields the MLE for σ^2

$$\hat{\sigma}^2_{\text{mle}} = \frac{SSE_n(\widehat{\boldsymbol{\beta}}_{\text{mle}})}{n} = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

which is the same as the moment estimator (3.28).

Plugging the estimates into (3.44) we obtain the maximized log-likelihood

$$\log L\left(\widehat{\boldsymbol{\beta}}_{\text{mle}}, \hat{\sigma}^2_{\text{mle}}\right) = -\frac{n}{2}\left(\log(2\pi) + 1\right) - \frac{n}{2}\log\left(\hat{\sigma}^2_{\text{mle}}\right). \qquad (3.45)$$

The log-likelihood (or the negative 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.

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) and cps09mar.txt (text format). The variables are:

1. age: years, capped at 85

2. female: 1 if female, 0 otherwise

3. hisp: 1 if Spanish, Hispanic, or Latino, 0 otherwise

4. education

0	Less than 1st grade
4	1st, 2nd, 3rd, or 4th grade
6	5th or 6th grade
8	7th or 8th grade
9	9th grade
10	10th grade
11	11th grade or 12th grade with no high school diploma
12	High school graduate, high school diploma or equivalent
13	Some college but no degree
14	Associate degree in college, including occupation/vocation programs
16	Bachelor's degree or equivalent (BA, AB, BS)
18	Master's degree (MA, MS MENG, MED, MSW, MBA)
20	Professional degree or Doctorate degree (MD, DDS, DVM, LLB, JD, PHD, EDD)

5. earnings: total annual wage and salary earnings

6. hours: number of hours worked per week

7. week: number of weeks worked per year

8. union: 1 for member of a labor union, 0 otherwise

9. uncov: 1 if covered by a union or employee association contract, 0 otherwise

10. region

1	Northeast
2	Midwest
3	South
4	West

11. Race

1	White only
2	Black only
3	American Indian, Alaskan Native (AI) only
4	Asian only
5	Hawaiian/Pacific Islander (HP) only
6	White-Black
7	White-AI
8	White-Asian
9	White-HP
10	Black-AI
11	Black-Asian
12	Black-HP
13	AI-Asian
14	Asian-HP
15	White-Black-AI
16	White-Black-Asian
17	White-AI-Asian
18	White-Asian-HP
19	White-Black-AI-Asian
20	2 or 3 races
21	4 or 5 races

12. marital

1	Married - civilian spouse present
2	Married - Armed Forces spouse present
3	Married - spouse absent (except separated)
4	Widowed
5	Divorced
6	Separated
7	Never married

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", while for Gauss and R there are no specific naming requirements.

To execute a program file, you type a command within the program.

Stata: `do chapter3` executes the file *chapter3.do*

Gauss: `run chapter3.prg` executes the file *chapter3.prg*

Matlab: `run chapter3` executes the file *chapter3.m*

R: `source(“chapter3.r”)` executes the file *chatper3.r*

When writing batch files, it is useful to include comments for documentation and readability.

We illustrate programming files for Stata, Gauss, R, and Matlab, which execute a portion of the empirical illustrations from Sections 3.7 and 3.17.

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
```

Gauss Program File

```
/* Load the data and create subsamples */
load dat[50742,12]=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=selif(dat,mbf);
dat2=selif(dat,sam);
/* First regression */
y=ln(dat1[.,5]./(dat1[.,6].*dat1[.,7]));
x=dat1[.,4]~ones(rows(dat1),1);
beta=invpd(x’x)*(x’y);
beta;
/* Second regression */
y=ln(dat2[.,5]./(dat2[.,6].*dat2[.,7]));
experience=dat2[.,1]-dat2[.,4]-6;
exp2 = (experience.^2)/100;
x=dat2[.,4]~experience~exp2~ones(rows(dat2),1);
beta=invpd(x’x)*(x’y);
beta;
/* Create leverage and influence */
e=y-x*beta;
leverage=sumc((x.*(x*invpd(x’x)))’);
d=leverage.*e./(1-leverage);
"Influence " maxc(abs(d));
```

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);
```

3.21 Technical Proofs*

Proof of Theorem 3.10.1, equation (3.23): 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.5))

$$\boldsymbol{P} = \boldsymbol{B}' \left[\begin{array}{cc} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{array} \right] \boldsymbol{B}$$

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

$$\begin{aligned} h_{ii} &= \boldsymbol{s}'\boldsymbol{B}' \left[\begin{array}{cc} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{array} \right] \boldsymbol{B}\boldsymbol{s} \\ &= \boldsymbol{b}_1' \left[\begin{array}{cc} \boldsymbol{I}_k & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{array} \right] \boldsymbol{b}_1 \\ &= \boldsymbol{b}_1'\boldsymbol{b}_1 \\ &\leq \boldsymbol{b}'\boldsymbol{b} \\ &= 1 \end{aligned}$$

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 (3.23). ■

Proof of Equation (3.37). The Sherman–Morrison formula (A.3) from Appendix A.5 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

$$\left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

and thus

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{(-i)} &= \left(\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i y_i \\ &+ (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{x}_i y_i\right) \\ &= \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i y_i + (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - h_{ii} y_i\right) \\ &= \widehat{\boldsymbol{\beta}} - (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \left((1 - h_{ii})\, y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + h_{ii} y_i\right) \\ &= \widehat{\boldsymbol{\beta}} - (1 - h_{ii})^{-1} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \hat{e}_i \end{aligned}$$

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 $\sigma^2 = \operatorname{var}(y)$. Define

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

Let $(\hat{\mu}, \hat{\sigma}^2)$ be the values such that $\overline{g}_n(\hat{\mu}, \hat{\sigma}^2) = \mathbf{0}$ where $\overline{g}_n(m, s) = n^{-1}\sum_{i=1}^n g\left(y_i, m, s\right)$. Show that $\hat{\mu}$ and $\hat{\sigma}^2$ 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 $\boldsymbol{X}'\hat{\boldsymbol{e}} = \mathbf{0}$.

Exercise 3.4 Let $\hat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$. Find $\boldsymbol{X}_2'\hat{\boldsymbol{e}}$.

Exercise 3.5 Let $\hat{\boldsymbol{e}}$ be the OLS residual from a regression of $\boldsymbol{y}$ on $\boldsymbol{X}$. Find the OLS coefficient from a regression of $\hat{\boldsymbol{e}}$ on $\boldsymbol{X}$.

Exercise 3.6 Let $\hat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$. Find the OLS coefficient from a regression of $\hat{\boldsymbol{y}}$ 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 $\operatorname{tr}\boldsymbol{M} = n - k$.

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, $\dfrac{1}{n}\sum_{i=1}^n \hat{y}_i = \overline{y}$.

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

$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.46}$$

$$\boldsymbol{y} = \boldsymbol{d}_1\alpha_1 + \boldsymbol{d}_2\alpha_2 + \boldsymbol{e} \tag{3.47}$$

$$\boldsymbol{y} = \mu + \boldsymbol{d}_1\phi + \boldsymbol{e} \tag{3.48}$$

Can all three equations (3.46), (3.47), and (3.48) be estimated by OLS? Explain if not.

(a) Compare regressions (3.47) and (3.48). Is one more general than the other? Explain the relationship between the parameters in (3.47) and (3.48).

(b) Compute $\boldsymbol{\iota}'\boldsymbol{d}_1$ and $\boldsymbol{\iota}'\boldsymbol{d}_2$, where $\boldsymbol{\iota}$ is an $n \times 1$ is a vector of ones.

(c) Letting $\boldsymbol{\alpha} = (\alpha_1\ \alpha_2)'$, write equation (3.47) 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

$$\boldsymbol{y} = \boldsymbol{d}_1 \hat{\gamma}_1 + \boldsymbol{d}_2 \hat{\gamma}_2 + \hat{\boldsymbol{u}},$$

show that $\hat{\gamma}_1$ is the sample mean of the dependent variable among the men of the sample ($\overline{y}_1$), and that $\hat{\gamma}_2$ 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 regresion

$$\boldsymbol{y}^* = \boldsymbol{X}^* \widetilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{e}}$$

with $\widehat{\boldsymbol{\beta}}$ from the OLS regression

$$\boldsymbol{y} = \boldsymbol{d}_1 \hat{\alpha}_1 + \boldsymbol{d}_2 \hat{\alpha}_2 + \boldsymbol{X} \widehat{\boldsymbol{\beta}} + \hat{\boldsymbol{e}}.$$

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 $\hat{\boldsymbol{y}}$.

Exercise 3.16 Consider two least-squares regressions

$$\boldsymbol{y} = \boldsymbol{X}_1 \widetilde{\boldsymbol{\beta}}_1 + \tilde{\boldsymbol{e}}$$

and

$$\boldsymbol{y} = \boldsymbol{X}_1 \widehat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2 \widehat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{e}}.$$

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 $\tilde{\sigma}^2 \geq \hat{\sigma}^2$. Is equality possible?

Exercise 3.18 For which observations will $\widehat{\boldsymbol{\beta}}_{(-i)} = \widehat{\boldsymbol{\beta}}$?

Exercise 3.19 Use the data set from Section 3.19 and the sub-sample used for equation (3.43) (see Section 3.20) for data construction)

1. Estimate equation (3.43) and compute the equation R^2 and sum of squared errors.

2. 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 (3.43)? Explain.

3. Do the R^2 and sum-of-squared errors from parts 1 and 2 equal? Explain.

Exercise 3.20 Estimate equation (3.43) as in part 1 of the previous question. Let $\hat{e}_i$ be the OLS residual, $\hat{y}_i$ the predicted value from the regression, x_{1i} be education and x_{2i} be experience. Numerically calculate the following:

(a) $\sum_{i=1}^n \hat{e}_i$

(b) $\sum_{i=1}^n x_{1i}\hat{e}_i$

(c) $\sum_{i=1}^n x_{2i}\hat{e}_i$

(d) $\sum_{i=1}^n x_{1i}^2\hat{e}_i$

(e) $\sum_{i=1}^n x_{2i}^2\hat{e}_i$

(f) $\sum_{i=1}^n \hat{y}_i\hat{e}_i$

(g) $\sum_{i=1}^n \hat{e}_i^2$

Are these calculations consistent with the theoretical properties of OLS? Explain.

Exercise 3.21 Use the data set from Section 3.19.

1. 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.

2. 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 least-squares applied to a random sample in the 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 Sample Mean

To start with the simplest setting, we first consider the intercept-only model

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

which is equivalent to the regression model with $k = 1$ and $x_i = 1$. In the intercept model, $\mu = \mathbb{E}\left(y_i\right)$ 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 (3.7).

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

$$\mathbb{E}\left(\overline{y}\right) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}y_i = \mu.$$

This shows that the expected value of 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 4.2.1 *An estimator $\widehat{\theta}$ for θ is* ***unbiased*** *if $\mathbb{E}\widehat{\theta} = \theta$.*

We next calculate the variance of the estimator $\overline{y}$. Making the substitution $y_i = \mu + e_i$ we find

$$\overline{y} - \mu = \frac{1}{n}\sum_{i=1}^{n} e_i.$$

Then

$$\begin{aligned}
\operatorname{var}(\overline{y}) &= \mathbb{E}(\overline{y}-\mu)^2 \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} e_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} e_j\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(e_i e_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\
&= \frac{1}{n}\sigma^2.
\end{aligned}$$

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.

4.3 Linear Regression Model

We now consider the linear regression model. Throughout the remainder of this chapter we maintain the following.

Assumption 4.3.1 ***Linear Regression Model***
The observations $(y_i, \boldsymbol{x}_i)$ *come from a random sample and satisfy the linear regression equation*

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \tag{4.1}$$

$$\mathbb{E}(e_i \mid \boldsymbol{x}_i) = 0. \tag{4.2}$$

The variables have finite second moments

$$\mathbb{E}y_i^2 < \infty,$$

$$\mathbb{E}\|\boldsymbol{x}_i\|^2 < \infty,$$

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 4.3.2 ***Homoskedastic Linear Regression Model***
In addition to Assumption 4.3.1,

$$\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2(\boldsymbol{x}_i) = \sigma^2 \tag{4.3}$$

is independent of $\boldsymbol{x}_i$.

4.4 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 (3.9), 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

$$\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \begin{pmatrix} \vdots \\ \mathbb{E}\left(e_i \mid \boldsymbol{X}\right) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) \\ \vdots \end{pmatrix} = \boldsymbol{0}. \tag{4.6}$$

Using definition (3.18), 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 (3.18) 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

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid \boldsymbol{X}\right) &= \mathbb{E}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) \\ &= \boldsymbol{0}.\end{aligned}$$

Regardless of the method, we have shown that $\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$. Applying the law of iterated expectations, we find that

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}}\right) = \mathbb{E}\left(\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)\right) = \boldsymbol{\beta}.$$

We have shown the following theorem.

Theorem 4.4.1 ***Mean of Least-Squares Estimator***
In the linear regression model (Assumption 4.3.1)

$$\mathbb{E}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta} \tag{4.8}$$

and

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}. \tag{4.9}$$

Equation (4.9) says that the estimator $\widehat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$, meaning that the distribution of $\widehat{\boldsymbol{\beta}}$ is centered at $\boldsymbol{\beta}$. Equation (4.8) says that the estimator is conditionally unbiased, which is a stronger result. It says that $\widehat{\boldsymbol{\beta}}$ is unbiased for any realization of the regressor matrix $\boldsymbol{X}$.

4.5 Variance of Least Squares Estimator

In this section we calculate the conditional variance of the OLS estimator.

For any $r \times 1$ random vector $\mathbf{Z}$ define the $r \times r$ covariance matrix

$$\begin{aligned}\operatorname{var}(\boldsymbol{Z}) &= \mathbb{E}\left(\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\right)\left(\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}\right)' \\ &= \mathbb{E}\boldsymbol{Z}\boldsymbol{Z}' - \left(\mathbb{E}\boldsymbol{Z}\right)\left(\mathbb{E}\boldsymbol{Z}\right)'\end{aligned}$$

and for any pair $(\mathbf{Z}, \mathbf{X})$ define the conditional covariance matrix

$$\operatorname{var}(\boldsymbol{Z} \mid \boldsymbol{X}) = \mathbb{E}\left((\boldsymbol{Z} - \mathbb{E}(\boldsymbol{Z} \mid \boldsymbol{X}))(\boldsymbol{Z} - \mathbb{E}(\boldsymbol{Z} \mid \boldsymbol{X}))' \mid \boldsymbol{X}\right).$$

We define

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} \overset{def}{=} \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$$

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

$$\boldsymbol{D} = \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right).$$

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 1.5.1) 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 4.5.1 ***Variance of Least-Squares Estimator***
In the linear regression model (Assumption 4.3.1)

$$\begin{aligned} \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} &= \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \qquad (4.12) \end{aligned}$$

where $\boldsymbol{D}$ *is defined in (4.10).*
In the homoskedastic linear regression model (Assumption 4.3.2)

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2.$$

4.6 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 4.6.1 ***Gauss-Markov***

1. *In the homoskedastic linear regression model (Assumption 4.3.2), the best (minimum-variance) unbiased linear estimator is the least-squares estimator*

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$$

2. *In the linear regression model (Assumption 4.3.1), the best unbiased linear estimator is*

$$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{y} \qquad (4.13)$$

The first part of the Gauss-Markov theorem is a limited 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.

The second part of the theorem shows that in the (heteroskedastic) linear regression model, within the class of linear unbiased estimators the best estimator is not least-squares but is (4.13). This is called the **Generalized Least Squares** (GLS) estimator. The GLS estimator is infeasible as the matrix $\boldsymbol{D}$ is unknown. This result does not suggest a practical alternative to least-squares. We return to the issue of feasible implementation of GLS in Section 9.2.

We give a proof of the first part of the theorem below, and leave the proof of the second part for Exercise 4.3.

Proof of Theorem 4.6.1.1. 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 $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2$ and that of $\boldsymbol{A}'\boldsymbol{y}$ is $\boldsymbol{A}'\boldsymbol{A}\sigma^2$. It is sufficient to show that the difference $\boldsymbol{A}'\boldsymbol{A} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ is positive semi-definite. Set $\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. Note that $\boldsymbol{X}'\boldsymbol{C} = \boldsymbol{0}$. Then we calculate that

$$\begin{aligned}
\boldsymbol{A}'\boldsymbol{A} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} &= \left(\boldsymbol{C} + \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)'\left(\boldsymbol{C} + \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{C}'\boldsymbol{C} + \boldsymbol{C}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{C} \\
&\quad + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{C}'\boldsymbol{C}.
\end{aligned}$$

The matrix $\boldsymbol{C}'\boldsymbol{C}$ is positive semi-definite (see Appendix A.8) as required.

4.7 Residuals

What are some properties of the residuals $\hat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ and prediction errors $\tilde{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)}$, at least in the context of the linear regression model?

Recall from (3.26) that we can write the residuals in vector notation as

$$\hat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$$

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

$$\mathbb{E}\left(\hat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \mathbb{E}\left(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{M}\mathbb{E}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{0}$$

and

$$\operatorname{var}\left(\hat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \operatorname{var}\left(\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}\right) = \boldsymbol{M}\,\mathbf{var}\left(\boldsymbol{e} \mid \boldsymbol{X}\right)\boldsymbol{M} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{M} \tag{4.14}$$

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 (4.14) simplies to

$$\operatorname{var}\left(\hat{\boldsymbol{e}} \mid \boldsymbol{X}\right) = \boldsymbol{M}\sigma^2. \tag{4.15}$$

In particular, for a single observation i, we can find the (conditional) variance of $\hat{e}_i$ by taking the i^{th} diagonal element of (4.16). Since the i^{th} diagonal element of $\boldsymbol{M}$ is $1 - h_{ii}$ as defined in (3.21) we obtain

$$\operatorname{var}(\hat{e}_i \mid \boldsymbol{X}) = \mathbb{E}\left(\hat{e}_i^2 \mid \boldsymbol{X}\right) = (1 - h_{ii})\,\sigma^2. \tag{4.16}$$

As this variance is a function of h_{ii} and hence $\boldsymbol{x}_i$, the residuals $\hat{e}_i$ are heteroskedastic even if the errors e_i are homoskedastic.

Similarly, recall from (3.40) that the prediction errors $\tilde{e}_i = (1 - h_{ii})^{-1}\,\hat{e}_i$ can be written in vector notation as $\tilde{\boldsymbol{e}} = \boldsymbol{M}^* \hat{\boldsymbol{e}}$ where $\boldsymbol{M}^*$ is a diagonal matrix with i^{th} diagonal element $(1 - h_{ii})^{-1}$. Thus $\tilde{\boldsymbol{e}} = \boldsymbol{M}^* \boldsymbol{M} \boldsymbol{e}$. We can calculate that

$$\mathbb{E}(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^* \boldsymbol{M} \mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = \boldsymbol{0}$$

and

$$\operatorname{var}(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}) = \boldsymbol{M}^* \boldsymbol{M} \,\mathbf{var}\,(\boldsymbol{e} \mid \boldsymbol{X})\, \boldsymbol{M} \boldsymbol{M}^* = \boldsymbol{M}^* \boldsymbol{M} \boldsymbol{D} \boldsymbol{M} \boldsymbol{M}^*$$

which simplifies under homoskedasticity to

$$\begin{aligned}\operatorname{var}(\tilde{\boldsymbol{e}} \mid \boldsymbol{X}) &= \boldsymbol{M}^* \boldsymbol{M} \boldsymbol{M} \boldsymbol{M}^* \sigma^2 \\ &= \boldsymbol{M}^* \boldsymbol{M} \boldsymbol{M}^* \sigma^2.\end{aligned}$$

The variance of the i^{th} prediction error is then

$$\begin{aligned}\operatorname{var}(\tilde{e}_i \mid \boldsymbol{X}) &= \mathbb{E}\left(\tilde{e}_i^2 \mid \boldsymbol{X}\right) \\ &= (1 - h_{ii})^{-1}(1 - h_{ii})(1 - h_{ii})^{-1}\sigma^2 \\ &= (1 - h_{ii})^{-1}\sigma^2.\end{aligned}$$

A residual with constant conditional variance can be obtained by rescaling. The **standardized residuals** are

$$\bar{e}_i = (1 - h_{ii})^{-1/2}\,\hat{e}_i, \tag{4.17}$$

and in vector notation

$$\bar{\boldsymbol{e}} = (\bar{e}_1, ..., \bar{e}_n)' = \boldsymbol{M}^{*1/2} \boldsymbol{M} \boldsymbol{e}.$$

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

$$\operatorname{var}(\bar{e}_i \mid \boldsymbol{X}) = \mathbb{E}\left(\bar{e}_i^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.18}$$

and thus these standardized residuals have the same bias and variance as the original errors when the latter are homoskedastic.

4.8 Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}e_i^2$ 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:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

and equals the MLE in the normal regression model (3.28).

In the linear regression model we can calculate the mean of $\hat{\sigma}^2$. From (3.26), the properties of projection matrices and the trace operator, observe that

$$\hat{\sigma}^2 = \frac{1}{n}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} = \frac{1}{n}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{M}\boldsymbol{e} = \frac{1}{n}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}'\right).$$

Then

$$\begin{aligned}\mathbb{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\right) \\ &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\right) \\ &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\boldsymbol{D}\right). \end{aligned} \tag{4.19}$$

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 (4.19) simplifies to

$$\begin{aligned}\mathbb{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\sigma^2\right) \\ &= \sigma^2\left(\frac{n-k}{n}\right),\end{aligned}$$

the final equality by (3.24). This calculation shows that $\hat{\sigma}^2$ 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 (4.16). Note that

$$\begin{aligned}\mathbb{E}\left(\hat{\sigma}^2 \mid \boldsymbol{X}\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\hat{e}_i^2 \mid \boldsymbol{X}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)\sigma^2 \\ &= \left(\frac{n-k}{n}\right)\sigma^2 \end{aligned} \tag{4.20}$$

the last equality using Theorem 3.10.1.

Since the bias takes a scale form, a classic method to obtain an unbiased estimator is by rescaling the estimator. Define

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{e}_i^2. \tag{4.21}$$

By the above calculation,

$$\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.22}$$

so

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

and 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 (4.17) is

$$\bar{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\bar{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}\hat{e}_i^2. \tag{4.23}$$

You can show (see Exercise 4.6) that

$$\mathbb{E}\left(\bar{\sigma}^2 \mid \boldsymbol{X}\right) = \sigma^2 \tag{4.24}$$

and thus $\bar{\sigma}^2$ 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 $\hat{\sigma}^2$, s^2 and $\bar{\sigma}^2$ are likely to be close. However, if not then s^2 and $\bar{\sigma}^2$ are generally preferred to $\hat{\sigma}^2$. Consequently it is best to use one of the bias-corrected variance estimators in applications.

4.9 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 $\boldsymbol{x}_{n+1}$ will be observed but not y_{n+1}. Given the coefficient estimate $\widehat{\boldsymbol{\beta}}$ the standard point estimate of $\mathbb{E}\left(y_{n+1} \mid \boldsymbol{x}_{n+1}\right) = \boldsymbol{x}_{n+1}'\boldsymbol{\beta}$ is $\tilde{y}_{n+1} = \boldsymbol{x}_{n+1}'\widehat{\boldsymbol{\beta}}$. The forecast error is the difference between the actual value y_{n+1} and the point forecast, $\tilde{e}_{n+1} = y_{n+1} - \tilde{y}_{n+1}$. The mean-squared forecast error (MSFE) is

$$MSFE_n = \mathbb{E}\tilde{e}_{n+1}^2.$$

In the linear regression model, $\tilde{e}_{n+1} = e_{n+1} - \boldsymbol{x}_{n+1}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$, so

$$\begin{aligned} MSFE_n &= \mathbb{E}e_{n+1}^2 - 2\mathbb{E}\left(e_{n+1}\boldsymbol{x}_{n+1}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right) \\ &\quad + \mathbb{E}\left(\boldsymbol{x}_{n+1}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{x}_{n+1}\right). \end{aligned} \tag{4.25}$$

The first term in (4.25) is σ^2. The second term in (4.25) 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 (4.25) is

$$\begin{aligned} &\operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\mathbb{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right) \\ &= \operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\mathbb{E}\left(\mathbb{E}\left(\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \mid \boldsymbol{X}\right)\right)\right) \\ &= \operatorname{tr}\left(\mathbb{E}\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\mathbb{E}\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\right) \\ &= \mathbb{E}\operatorname{tr}\left(\left(\boldsymbol{x}_{n+1}\boldsymbol{x}_{n+1}'\right)\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\right) \\ &= \mathbb{E}\left(\boldsymbol{x}_{n+1}'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}_{n+1}\right) \end{aligned} \tag{4.26}$$

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 (3.41)

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2$$

where $\tilde{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{(-i)} = \hat{e}_i(1-h_{ii})^{-1}$. Indeed, we can calculate that

$$\begin{aligned}\mathbb{E}\tilde{\sigma}^2 &= \mathbb{E}\tilde{e}_i^2 \\ &= \mathbb{E}\left(e_i - \boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)\right)^2 \\ &= \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}}_{(-i)} - \boldsymbol{\beta}\right)'\boldsymbol{x}_i\right).\end{aligned}$$

By a similar calculation as in (4.26) we find

$$\mathbb{E}\tilde{\sigma}^2 = \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_i'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}_{(-i)}}\boldsymbol{x}_i\right) = MSFE_{n-1}.$$

his is the MSFE based on a sample of size $n-1$, rather than size n. The difference arises because the in-sample prediction errors $\tilde{e}_i$ for $i \le n$ are calculated using an effective sample size of $n-1$, while the out-of sample prediction error $\tilde{e}_{n+1}$ 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 $\tilde{\sigma}^2$ is a reasonable estimator for $MSFE_n$.

Theorem 4.9.1 ***MSFE***
In the linear regression model (Assumption 4.3.1)

$$MSFE_n = \mathbb{E}\tilde{e}_{n+1}^2 = \sigma^2 + \mathbb{E}\left(\boldsymbol{x}_{n+1}'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}_{n+1}\right)$$

where $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$. *Furthermore,* $\tilde{\sigma}^2$ *defined in (3.41) is an unbiased estimator of* $MSFE_{n-1}$:

$$\mathbb{E}\tilde{\sigma}^2 = MSFE_{n-1}$$

4.10 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 4.3.2).

Under homoskedasticity, the covariance matrix takes the relatively simple form

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2$$

which is known up to the unknown scale σ^2. In Section 4.8 we discussed three estimators of σ^2. The most commonly used choice is s^2, leading to the classic covariance matrix estimator

$$\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}s^2. \tag{4.27}$$

Since s^2 is conditionally unbiased for σ^2, it is simple to calculate that $\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}}$ is conditionally unbiased for $\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}$ under the assumption of homoskedasticity:

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\ &= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}.\end{aligned}$$

This estimator was the dominant covariance matrix estimator in applied econometrics for many years, and is still the default method in most regression packages.

If the estimator (4.27) is used, but the regression error is heteroskedastic, it is possible for $\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}}$ 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

$$\frac{\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}}{\mathbb{E}\left(\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right)} = \frac{\sum_{i=1}^n x_i^4}{\sigma^2 \sum_{i=1}^n x_i^2} \simeq \frac{\mathbb{E}x_i^4}{\left(\mathbb{E}x_i^2\right)^2} = \kappa.$$

(Notice that we use the fact that $\sigma_i^2 = x_i^2$ implies $\sigma^2 = \mathbb{E}\sigma_i^2 = \mathbb{E}x_i^2$.) The constant κ is the standardized forth 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 $\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}}$. 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 $\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\beta}}}$. While this is an extreme and constructed example, the point is that the classic covariance matrix estimator (4.27) may be quite biased when the homoskedasticity assumption fails.

4.11 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 contruct 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

$$\begin{aligned}\boldsymbol{D} &= \mathrm{diag}\left(\sigma_1^2, ..., \sigma_n^2\right)\\ &= \mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right)\\ &= \mathbb{E}\left(\boldsymbol{D}_0 \mid \boldsymbol{X}\right)\end{aligned}$$

where $\boldsymbol{D}_0 = \mathrm{diag}\left(e_1^2, ..., e_n^2\right)$. 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

$$\begin{aligned}\widehat{\boldsymbol{V}}^{ideal}_{\widehat{\boldsymbol{\beta}}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}_0\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' e_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned}$$

Indeed,

$$\begin{aligned}\mathbb{E}\left(\widehat{\boldsymbol{V}}^{ideal}_{\widehat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\mathbb{E}\left(e_i^2 \mid \boldsymbol{X}\right)\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\sigma_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\\ &= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\end{aligned}$$

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 $\hat{e}_i$. Making this substitution we obtain the estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.28}$$

We know, however, that $\hat{e}_i^2$ is biased towards zero. To estimate the variance σ^2 the unbiased estimator s^2 scales the moment estimator $\hat{\sigma}^2$ by $n/(n-k)$. Making the same adjustment we obtain the estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \left(\frac{n}{n-k}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.29}$$

While the scaling by $n/(n-k)$ is ad hoc, it is recommended over the unscaled estimator (4.28).

Alternatively, we could use the prediction errors $\tilde{e}_i$ or the standardized residuals $\bar{e}_i$, yielding the estimators

$$\begin{aligned}\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\tilde{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\end{aligned} \tag{4.30}$$

and

$$\begin{aligned}\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\bar{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\left(1-h_{ii}\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned} \tag{4.31}$$

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 $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ 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 scaled estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is the default robust covariance matrix estimator implemented in Stata. The estimator $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ was introduced by Andrews (1991) based on the principle of leave-one-out cross-validation (and is implemented using the vce(hc3) option in Stata). 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).

Since $(1-h_{ii})^{-2} > (1-h_{ii})^{-1} > 1$ it is straightforward to show that

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} < \overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} < \widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \tag{4.32}$$

(See Exercise 4.7). 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 (4.16),

$$\begin{aligned}
\mathbb{E}\left(\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} \mid \boldsymbol{X}\right) &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\mathbb{E}\left(\hat{e}_i^2 \mid \boldsymbol{X}\right)\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\left(1-h_{ii}\right)\sigma^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'h_{ii}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\
&< \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\
&= \boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}.
\end{aligned}$$

This calculation shows that $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ is biased towards zero.

Similarly, (again under homoskedasticity) we can calculate that $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is biased away from zero, specifically

$$\mathbb{E}\left(\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) > \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \tag{4.33}$$

while the estimator $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is unbiased

$$\mathbb{E}\left(\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2. \tag{4.34}$$

(See Exercise 4.8.)

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.

We have introduced five covariance matrix estimators, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, $\widetilde{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$, and $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$. Which should you use? The classic estimator $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$ 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 $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0}$, so users must intentionally select a robust covariance matrix estimator.

Of the four robust estimators, $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W}$ and $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ are the most commonly used, and in particular $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is the default robust covariance matrix option in Stata. However, $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ may be the preferred choice based on its improved bias. As $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ 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.

4.12 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 4.12.1 *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,

$$s(\hat{\beta}_j) = \sqrt{\widehat{\boldsymbol{V}}_{\hat{\beta}_j}} = \sqrt{\left[\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\right]_{jj}}.$$

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 (3.11) 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

$$\begin{aligned} \overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} &= \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \begin{pmatrix} 763.26 & 48.513 \\ 48.513 & 3.1078 \end{pmatrix} \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.001 & -0.015 \\ -0.015 & 0.243 \end{pmatrix}. \end{aligned} \tag{4.35}$$

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

$$\begin{array}{rcccc} \widehat{\log(Wage)} = & \underset{(0.031)}{0.155} & Education + & \underset{(0.493)}{0.698} & . \end{array}$$

Alternatively, standard errors could be calculated using the other formulae. We report the different standard errors in the following table.

	Education	Intercept
Homoskedastic (4.27)	0.045	0.707
White (4.28)	0.029	0.461
Scaled White (4.29)	0.030	0.486
Andrews (4.30)	0.033	0.527
Horn-Horn-Duncan (4.31)	0.031	0.493

The homoskedastic standard errors are noticably different (larger, in this case) than the others, but the four robust standard errors are quite close to one another.

4.13 Computation

We illustrate methods to compute standard errors for equation (3.12) extending the code of Section 3.20.

Stata do File (continued)

```
*       Homoskedastic formula (4.27):
reg wage education experience exp2 if (mnwf == 1)
*       Scaled White formula (4.29):
reg wage education experience exp2 if (mnwf == 1), r
*       Andrews formula (4.30):
reg wage education experience exp2 if (mnwf == 1), vce(hc3)
*       Horn-Horn-Duncan formula (4.31):
reg wage education experience exp2 if (mnwf == 1), vce(hc2)
```

Gauss Program File (continued)

```
n=rows(y);
k=cols(x);
a=n/(n-k);
sig2=(e’e)/(n-k);
u1=x.*e;
u2=x.*(e./(1-leverage));
u3=x.*(e./sqrt(1-leverage));
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 @
```

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
```

4.14 Measures of Fit

As we described in the previous chapter, a commonly reported measure of regression fit is the regression R^2 defined as

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}.$$

where $\hat{\sigma}_y^2 = n^{-1}\sum_{i=1}^n (y_i - \overline{y})^2$. R^2 can be viewed as an estimator of the population parameter

$$\rho^2 = \frac{\operatorname{var}(\boldsymbol{x}_i'\boldsymbol{\beta})}{\operatorname{var}(y_i)} = 1 - \frac{\sigma^2}{\sigma_y^2}.$$

However, $\hat{\sigma}^2$ and $\hat{\sigma}_y^2$ are biased estimators. Theil (1961) proposed replacing these by the unbiased versions s^2 and $\tilde{\sigma}_y^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \overline{y})^2$ yielding what is known as **R-bar-squared** or **adjusted R-squared**:

$$\overline{R}^2 = 1 - \frac{s^2}{\tilde{\sigma}_y^2} = 1 - \frac{(n-1)\sum_{i=1}^n \hat{e}_i^2}{(n-k)\sum_{i=1}^n (y_i - \bar{y})^2}.$$

While $\overline{R}^2$ is an improvement on R^2, a much better improvement is

$$\widetilde{R}^2 = 1 - \frac{\sum_{i=1}^n \tilde{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = 1 - \frac{\tilde{\sigma}^2}{\hat{\sigma}_y^2}$$

where $\tilde{e}_i$ are the prediction errors (3.38) and $\tilde{\sigma}^2$ is the MSPE from (3.41). As described in Section (4.9), $\tilde{\sigma}^2$ 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 $\tilde{\sigma}^2$ 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 $\tilde{\sigma}^2$ to compare and select models. Models with high $\widetilde{R}^2$ (or low $\tilde{\sigma}^2$) 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 $\tilde{\sigma}^2$ in regression analysis, and omit R^2 and $\overline{R}^2$.

Henri Theil

Henri Theil (1924-2000) of Holland 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).

4.15 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)

	$\hat{\beta}$	$s(\hat{\beta})$
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
$\hat{\sigma}$	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 apparant 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.

4.16 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 $\boldsymbol{\alpha} \neq \mathbf{0}$ such that $\boldsymbol{X}\boldsymbol{\alpha} = \mathbf{0}$. 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 $\log(p_1/p_2)$, for 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 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 $\bar{y}/\hat{\sigma}_{\bar{y}}$ 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.

4.17 Normal Regression Model

In the special case of the normal linear regression model introduced in Section 3.18, we can derive exact sampling distributions for the least-squares estimator, residuals, and variance estimator.

In particular, under the normality assumption $e_i \mid \boldsymbol{x}_i \sim \mathrm{N}\left(0, \sigma^2\right)$ then we have 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. Since linear functions of normals are also normal, this implies that conditional on $\boldsymbol{X}$

$$\left(\begin{array}{c} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\boldsymbol{e}} \end{array}\right) = \left(\begin{array}{c} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}' \\ \boldsymbol{M} \end{array}\right) \boldsymbol{e} \sim \mathrm{N}\left(0, \left(\begin{array}{cc} \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} & 0 \\ 0 & \sigma^2 \boldsymbol{M} \end{array}\right)\right)$$

where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. Since uncorrelated normal variables are independent, it follows that $\widehat{\boldsymbol{\beta}}$ is independent of any function of the OLS residuals including the estimated error variance s^2 or $\hat{\sigma}^2$ or prediction errors $\tilde{\boldsymbol{e}}$.

The spectral decomposition (see equation (A.5)) of $\boldsymbol{M}$ yields

$$\boldsymbol{M} = \boldsymbol{H} \left[\begin{array}{cc} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{array}\right] \boldsymbol{H}'$$

where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$. Let $\boldsymbol{u} = \sigma^{-1}\boldsymbol{H}'\boldsymbol{e} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{H}'\boldsymbol{H}\right) \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$. Then

$$\begin{aligned}
\frac{n\hat{\sigma}^2}{\sigma^2} &= \frac{(n-k)\,s^2}{\sigma^2} \\
&= \frac{1}{\sigma^2}\hat{\boldsymbol{e}}'\hat{\boldsymbol{e}} \\
&= \frac{1}{\sigma^2}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} \\
&= \frac{1}{\sigma^2}\boldsymbol{e}'\boldsymbol{H}\begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{H}'\boldsymbol{e} \\
&= \boldsymbol{u}'\begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{u} \\
&\sim \chi^2_{n-k},
\end{aligned}$$

a chi-square distribution with $n-k$ degrees of freedom.

Furthermore, if standard errors are calculated using the homoskedastic formula (4.27)

$$\frac{\widehat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} = \frac{\widehat{\beta}_j - \beta_j}{s\sqrt{\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}}} \sim \frac{\mathrm{N}\left(0, \sigma^2\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}\right)}{\sqrt{\frac{\sigma^2}{n-k}\chi^2_{n-k}}\sqrt{\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}}} = \frac{\mathrm{N}\left(0,1\right)}{\sqrt{\frac{\chi^2_{n-k}}{n-k}}} \sim t_{n-k}$$

a t distribution with $n-k$ degrees of freedom.

Theorem 4.17.1 ***Normal Regression***
In the linear regression model (Assumption 4.3.1) if e_i is independent of $\boldsymbol{x}_i$ and distributed $\mathrm{N}\left(0, \sigma^2\right)$ *then*

- $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \sim \mathrm{N}\left(\boldsymbol{0}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$
- $\frac{n\hat{\sigma}^2}{\sigma^2} = \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$
- $\frac{\hat{\beta}_j - \beta_j}{s(\hat{\beta}_j)} \sim t_{n-k}$

These are the exact finite-sample distributions of the least-squares estimator and variance estimators, and are the basis for traditional inference in linear regression.

While elegant, the difficulty in applying Theorem 4.17.1 is that the normality assumption is too restrictive to be empirically plausible, and therefore inference based on Theorem 4.17.1 has no guarantee of accuracy. We develop an alternative inference theory based on large sample (asymptotic) approximations in the following chapter.

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's t rather than Gosset's t!

Exercises

Exercise 4.1 Explain the difference between $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'$ and $\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)$.

Exercise 4.2 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 $\hat{e}_i$ is the OLS residual from the regression of y_i on x_i, then $\sum_{i=1}^{n} x_i^2\hat{e}_i = 0$.

Exercise 4.3 Prove Theorem 4.6.1.2.

Exercise 4.4 In a linear model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}, \quad \mathbb{E}(\boldsymbol{e} \mid \boldsymbol{X}) = 0, \quad \text{var}\left(\boldsymbol{e} \mid \boldsymbol{X}\right) = \sigma^2\boldsymbol{\Omega}$$

with $\boldsymbol{\Omega}$ a known function of $\boldsymbol{X}$, the GLS estimator is

$$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{y}\right),$$

the residual vector is $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}$, and an estimate of σ^2 is

$$s^2 = \frac{1}{n-k}\hat{\boldsymbol{e}}'\boldsymbol{\Omega}^{-1}\hat{\boldsymbol{e}}.$$

(a) Find $\mathbb{E}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$.

(b) Find $\text{var}\left(\widetilde{\boldsymbol{\beta}} \mid \boldsymbol{X}\right)$.

(c) Prove that $\hat{\boldsymbol{e}} = \boldsymbol{M}_1\boldsymbol{e}$, where $\boldsymbol{M}_1 = \boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}$.

(d) Prove that $\boldsymbol{M}_1'\boldsymbol{\Omega}^{-1}\boldsymbol{M}_1 = \boldsymbol{\Omega}^{-1} - \boldsymbol{\Omega}^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}^{-1}$.

(e) Find $\mathbb{E}\left(s^2 \mid \boldsymbol{X}\right)$.

(f) Is s^2 a reasonable estimator for σ^2?

Exercise 4.5 Let $(y_i, \boldsymbol{x}_i)$ be a random sample with $\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{X}) = \boldsymbol{X\beta}$. Consider the **Weighted Least Squares** (WLS) estimator of $\boldsymbol{\beta}$

$$\widetilde{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}\right)$$

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 $\widetilde{\boldsymbol{\beta}}$ be a good estimator?

(b) Using your intuition, in which situations would you expect that $\widetilde{\boldsymbol{\beta}}$ would perform better than OLS?

Exercise 4.6 Show (4.24) in the homoskedastic regression model.

Exercise 4.7 Prove (4.32).

Exercise 4.8 Show (4.33) and (4.34) in the homoskedastic regression model.

Exercise 4.9 Let $\mu = \mathbb{E}\left(y_i\right), \sigma^2 = \mathbb{E}\left(y_i - \mu\right)^2$ and $\mu_3 = \mathbb{E}\left(y_i - \mu\right)^3$ and consider the sample mean $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Find $\mathbb{E}\left(\overline{y} - \mu\right)^3$ as a function of μ, σ^2, μ_3 and n.

Exercise 4.10 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 4.11 Continue the empirical analysis in Exercise 3.19.

1. Calculate standard errors using the homoskedasticity formula and using the four covariance matrices from Section 4.11.
2. Repeat in your second programming language. Are they identical?

Exercise 4.12 Continue the empirical analysis in Exercise 3.21. Calculate standard errors using the Horn-Horn-Duncan method. Repeat in your second programming language. Are they identical?

Chapter 5

An Introduction to Large Sample Asymptotics

5.1 Introduction

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, nor sufficient for inference (confidence intervals and hypothesis testing) on the unknown parameters. Furthermore, the theory does not apply in the context of the linear projection model, which is more relevant for empirical applications.

To illustrate the situation with an example, let y_i and x_i be drawn from the joint density

$$f(x,y) = \frac{1}{2\pi xy} \exp\left(-\frac{1}{2}(\log y - \log x)^2\right) \exp\left(-\frac{1}{2}(\log x)^2\right)$$

and let $\hat{\beta}$ 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 $\hat{\beta}$ was computed and plotted in Figure 5.1 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 $\hat{\beta}$?

In principle the sampling distribution of $\hat{\beta}$ 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 $\hat{\beta}$ except in very special cases. Therefore we typically rely on approximation methods.

The most widely used and versatile method is asymptotic theory, which approximates sampling distributions 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.

In this chapter we provide a concise summary. It will be useful for most students to review this material, even if most is familiar.

5.2 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

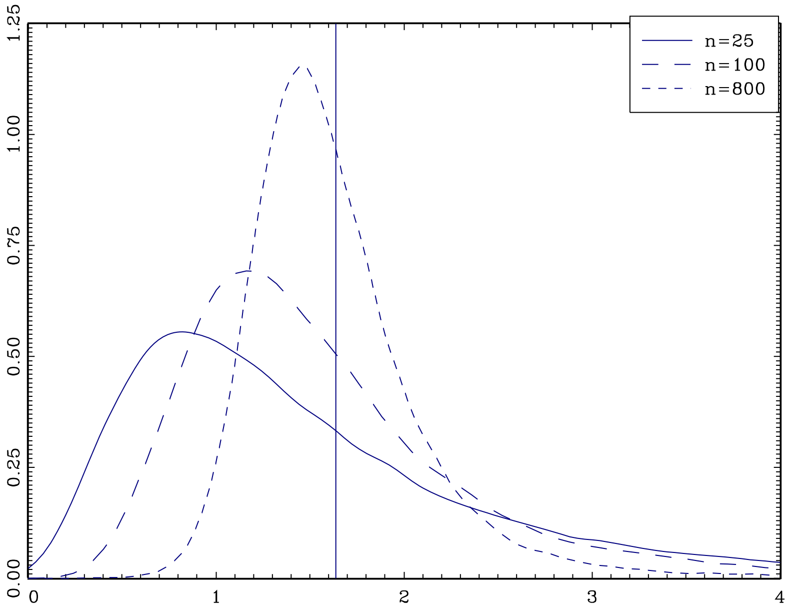


Figure 5.1: Sampling Density of $\hat{\beta}$

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 5.2.1 *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 sequence

Definition 5.2.2 $\liminf_{n\to\infty} a_n \overset{def}{=} \lim_{n\to\infty} \inf_{m\geq n} a_n$

Definition 5.2.3 $\limsup_{n\to\infty} a_n \overset{def}{=} \lim_{n\to\infty} \sup_{m\geq n} a_n$

The limit inferior and limit superior always exist, 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.

5.3 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 5.3.1 *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* $\delta > 0$,

$$\lim_{n\to\infty} \Pr\left(|z_n - z| \le \delta\right) = 1. \tag{5.1}$$

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 $\{|z_n - z| \le \delta\}$ occurs when z_n is within δ of the point z. $\Pr\left(|z_n - z| \le \delta\right)$ is the probability of this event – that z_n is within δ of the point z. Equation (5.1) 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 to confuse convergence in probability with convergence in expectation:

$$\mathbb{E}z_n \longrightarrow \mathbb{E}z. \tag{5.2}$$

They are related but distinct concepts. Neither (5.1) nor (5.2) 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

$$\begin{aligned} \Pr\left(z_n = 0\right) &= 1 - \frac{1}{n} \\ \Pr\left(z_n = a_n\right) &= \frac{1}{n}. \end{aligned} \tag{5.3}$$

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.

5.4 Weak Law of Large Numbers

In large samples we expect parameter estimates to be close to the population values. For example, in Section 4.2 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 5.4.1 ***Chebyshev's Inequality**. For any random variable z_n and constant $\delta > 0$*

$$\Pr\left(|z_n - \mathbb{E}z_n| > \delta\right) \leq \frac{\mathrm{var}(z_n)}{\delta^2}.$$

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

$$\Pr\left(|z_n - \mathbb{E}z_n| > \delta\right) = \Pr\left((z_n - \mathbb{E}z_n)^2 > \delta^2\right) = \int_{\{u^2>\delta^2\}} dF_n(u).$$

The integral is over the event $\left\{u^2 > \delta^2\right\}$, so that the inequality $1 \leq \dfrac{u^2}{\delta^2}$ holds throughout. Thus

$$\int_{\{u^2>\delta^2\}} dF_n(u) \leq \int_{\{u^2>\delta^2\}} \frac{u^2}{\delta^2} dF_n(u) \leq \int \frac{u^2}{\delta^2} dF_n(u) = \frac{\mathbb{E}\left(z_n - \mathbb{E}z_n\right)^2}{\delta^2} = \frac{\mathrm{var}(z_n)}{\delta^2},$$

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 $\delta > 0$

$$\Pr\left(|\overline{y} - \mathbb{E}\overline{y}| > \delta\right) \leq \frac{\sigma^2/n}{\delta^2}.$$

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 ε.) Thus the probability that $\overline{y}$ is within δ of $\mathbb{E}\overline{y} = \mu$ approaches 1 as n gets large, or

$$\lim_{n\to\infty} \Pr\left(|\overline{y} - \mu| \leq \delta\right) = 1.$$

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 proof all that is necessary is a finite mean.

Theorem 5.4.2 ***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 5.4.2 is presented in Section 5.14.

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 5.4.1 *An estimator $\hat{\theta}$ of a parameter θ is* ***consistent*** *if $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$.*

Theorem 5.4.3 *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 $\hat{\theta}$ 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.

5.5 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 5.5.1 *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 $\delta > 0$*

$$\Pr\left(\lim_{n\to\infty} |z_n - z| \leq \delta\right) = 1. \tag{5.4}$$

The convergence (5.4) is stronger than (5.1) 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 (5.3) of Section 5.3, 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 5.5.1 ***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, Section 22) 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.

5.6 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 5.6.1 *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 5.6.2 ***Weak Law of Large Numbers (WLLN) for random vectors***
If $\boldsymbol{y}_i$ are independent and identically distributed and $\mathbb{E}\|\boldsymbol{y}\| < \infty$, then as $n \to \infty$,

$$\overline{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \xrightarrow{p} \mathbb{E}(\boldsymbol{y}).$$

5.7 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.

Definition 5.7.1 *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 $n \to \infty$.*

When $\boldsymbol{z}_n \xrightarrow{d} \boldsymbol{z}$, it is common to refer to $\boldsymbol{z}$ as the **asymptotic 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, $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{c}$.

The typical path to establishing convergence in distribution is through the central limit theorem (CLT), which states that a standardized sample average converges in distribution to a normal random vector.

Theorem 5.7.1 ***Lindeberg–Lévy Central Limit Theorem (CLT).*** *If $\boldsymbol{y}_i$ are independent and identically distributed and $\mathbb{E}\|\boldsymbol{y}\|^2 < \infty$, then as $n \to \infty$*

$$\sqrt{n}\left(\overline{\boldsymbol{y}} - \boldsymbol{\mu}\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\boldsymbol{y}_i - \boldsymbol{\mu}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}\right)$$

where $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{y}$ and $\boldsymbol{V} = \mathbb{E}\left(\left(\boldsymbol{y} - \boldsymbol{\mu}\right)\left(\boldsymbol{y} - \boldsymbol{\mu}\right)'\right)$.

The standardized sum $\boldsymbol{z}_n = \sqrt{n}\left(\overline{\boldsymbol{y}}_n - \boldsymbol{\mu}\right)$ has mean zero and variance $\boldsymbol{V}$. What the CLT adds is that the variable $\boldsymbol{z}_n$ 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 does not require the restriction to identical distributions was provided by Lindeberg (1922).

Theorem 5.7.2 ***Lindeberg Central Limit Theorem (CLT).*** *Suppose that y_i are independent but not necessarily identically distributed with finite means $\mu_i = \mathbb{E}y_i$ and variances $\sigma_i^2 = \mathbb{E}\left(y_i - \mu_i\right)^2$. Set $\nu_n^2 = \sum_{i=1}^n \sigma_i^2$. If for all $\varepsilon > 0$*

$$\lim_{n\to\infty} \frac{1}{\nu_n^2} \sum_{i=1}^{n} \mathbb{E}\left(y_i - \mu_i\right)^2 1\left(\left|y_i - \mu_i\right| \geq \varepsilon \nu_n\right) = 0 \tag{5.5}$$

then

$$\frac{1}{\nu_n} \sum_{i=1}^{n} \left(y_i - \mu_i\right) \xrightarrow{d} \mathrm{N}\left(0, 1\right).$$

Equation (5.5) is known as Lindeberg's condition. A standard method to verify (5.5) is via Lyapunov's condition: For some $\delta > 0$

$$\lim_{n\to\infty} \frac{1}{\nu_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left(y_i - \mu_i\right)^{2+\delta} = 0. \tag{5.6}$$

It is easy to verify that (5.6) implies (5.5), and (5.6) is often easy to verify. For example, if $\sup_i \mathbb{E}\left(y_i - \mu_i\right)^3 \leq \kappa < \infty$ and $\inf_i \sigma_i^2 \geq c > 0$ then

$$\frac{1}{\nu_n^3} \sum_{i=1}^{n} \mathbb{E}\left(y_i - \mu_i\right)^3 \leq \frac{n\kappa}{(nc)^{3/2}} \to 0$$

so (5.6) is satisfied.

5.8 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

$$\boldsymbol{\mu} = \mathbb{E}\boldsymbol{h}(\boldsymbol{y})$$

for some function $\boldsymbol{h} : \mathbb{R}^m \to \mathbb{R}^k$. For example, the second moment of y is $\mathbb{E}y^2$, the k'th is $\mathbb{E}y^k$, the moment generating function is $\mathbb{E}\exp(ty)$, and the distribution function is $\mathbb{E}1\{y \leq x\}$.

Estimating parameters of this form fits into our previous analysis by defining the random variable $\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{y})$ for then $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{z}$ is just a simple moment of $\boldsymbol{z}$. This suggests the moment estimator

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

For example, the moment estimator of $\mathbb{E}y^k$ is $n^{-1}\sum_{i=1}^n y_i^k$, that of the moment generating function is $n^{-1}\sum_{i=1}^n \exp(ty_i)$, and for the distribution function the estimator is $n^{-1}\sum_{i=1}^n 1\{y_i \leq x\}$

Since $\widehat{\boldsymbol{\mu}}$ is a sample average, and transformations of iid variables are also iid, the asymptotic results of the previous sections immediately apply.

Theorem 5.8.1 *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 $n \to \infty$, $\widehat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$.*

Theorem 5.8.2 *If $\boldsymbol{y}_i$ are independent and identically distributed, $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{h}(\boldsymbol{y})$, and $\mathbb{E}\|\boldsymbol{h}(\boldsymbol{y})\|^2 < \infty$, then for $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{h}(\boldsymbol{y}_i)$, as $n \to \infty$,*

$$\sqrt{n}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V})$$

where $\boldsymbol{V} = \mathbb{E}\left((\boldsymbol{h}(\boldsymbol{y}) - \boldsymbol{\mu})(\boldsymbol{h}(\boldsymbol{y}) - \boldsymbol{\mu})'\right)$.

Theorems 5.8.1 and 5.8.2 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 5.8.1 and 5.8.2 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 momnets 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 $\mathrm{var}(\widehat{\mu}_8) = n^{-1}2,016,000$, 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 $\mathrm{var}(\widehat{\mu}_8) = n^{-1}\left(\mathbb{E}y^{16} - \left(\mathbb{E}y^8\right)^2\right)$. Since y is $\mathrm{N}(0,1)$, $\mathbb{E}y^{16} = 15!! = 2,027,025$ and $\mathbb{E}y^8 = 7!! = 105$ where $k!! = k(k-2)\cdots$ is the double factorial for odd k.

5.9 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

$$\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\mu}) = \boldsymbol{g}(\mathbb{E}\boldsymbol{h}(\boldsymbol{y})) \tag{5.7}$$

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

$$\gamma = \exp\left(\mathbb{E}\left(\log(w)\right)\right). \tag{5.8}$$

This is (5.7) 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 (5.7) 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

$$sk = \frac{\mathbb{E}(w - \mathbb{E}w)^3}{\left(\mathbb{E}(w - \mathbb{E}w)^2\right)^{3/2}}.$$

This is (5.7) with

$$\boldsymbol{h}(w) = \begin{pmatrix} w \\ w^2 \\ w^3 \end{pmatrix}$$

and

$$g(\mu_1, \mu_2, \mu_3) = \frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{\left(\mu_2 - \mu_1^2\right)^{3/2}}. \tag{5.9}$$

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 (5.8) 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

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n w_i^2 - \left(\frac{1}{n}\sum_{i=1}^n w_i\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (w_i - \overline{w})^2 .\end{aligned}$$

Thye 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 5.9.1 ***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 5.9.1 is given in Section 5.14.

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 $\mathbb{E}\left\|\boldsymbol{h}(\boldsymbol{y})\right\| < \infty$, 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 5.9.2 *If $\boldsymbol{y}_i$ are independent and identically distributed, $\boldsymbol{\beta} = \boldsymbol{g}(\mathbb{E}\boldsymbol{h}(\boldsymbol{y}))$, $\mathbb{E}\left\|\boldsymbol{h}(\boldsymbol{y})\right\| < \infty$, 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 5.9.2 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 5.6.2 and Theorem 5.9.2 that if $\mathbb{E}\left|\log(wage)\right| < \infty$ 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 $\mathbb{E}w^2 < \infty$ then as $n \to \infty$, $\widehat{\sigma}^2 \xrightarrow{p} \sigma^2$.

In our third example g defined in (5.9) 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 $\mathbb{E}\left|w\right|^3 < \infty$ and $\text{var}(w) > 0$ then as $n \to \infty$, $\widehat{sk} \xrightarrow{p} sk$.

5.10 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 5.10.1 ***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 5.10.1 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 5.10.1 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 5.10.2 ***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 5.10.1 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 5.10.3 ***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* $n \to \infty$

$$\sqrt{n}\left(\boldsymbol{g}(\widehat{\boldsymbol{\mu}}) - \boldsymbol{g}(\boldsymbol{\mu})\right) \xrightarrow{d} \boldsymbol{G}'\boldsymbol{\xi} \tag{5.10}$$

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* $n \to \infty$

$$\sqrt{n}\left(\boldsymbol{g}(\widehat{\boldsymbol{\mu}}) - \boldsymbol{g}(\boldsymbol{\mu})\right) \xrightarrow{d} \mathrm{N}\left(0, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}\right). \tag{5.11}$$

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 5.8.2 and 5.10.3 we can find the asymptotic distribution of the plug-in estimator $\widehat{\boldsymbol{\beta}}$.

Theorem 5.10.4 *If $\boldsymbol{y}_i$ are independent and identically distributed, $\boldsymbol{\mu} = \mathbb{E}\boldsymbol{h}(\boldsymbol{y})$, $\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\mu})$, $\mathbb{E}\left\|\boldsymbol{h}(\boldsymbol{y})\right\|^2 < \infty$, 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 $n \to \infty$*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}\right)$$

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 5.9.2 established the consistency of $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, and Theorem 5.10.4 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 asymptotic normality required that $\boldsymbol{g}(\boldsymbol{u})$ be continuously differentiable.

5.11 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 $n \to \infty$. We also write

$$z_n = o_p(a_n)$$

if $a_n^{-1}z_n = o_p(1)$. For example, for any consistent estimator $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ we can write

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_p(1).$$

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

$$\limsup_{n\to\infty} \Pr\left(|z_n| > M_\varepsilon\right) \le \varepsilon.$$

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 $\boldsymbol{z} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{V})$) then $\boldsymbol{z}_n = O_p(1)$. It follows that for estimators $\widehat{\boldsymbol{\beta}}$ which satisfy the convergence of Theorem 5.10.4 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 shrinking to zero at the rate $n^{-1/2}$.

Another useful observation is that a random sequence with a bounded moment is stochastically bounded.

Theorem 5.11.1 *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.21). The assumptions imply that there is some $M < \infty$ such that $\mathbb{E}\left\|\boldsymbol{z}_n\right\|^\delta \le Ma_n$ for all n. For any ε set $B = \left(\dfrac{M}{\varepsilon}\right)^{1/\delta}$. Then

$$\Pr\left(a_n^{-1/\delta}\left\|\boldsymbol{z}_n\right\| > B\right) = \Pr\left(\left\|\boldsymbol{z}_n\right\|^\delta > \frac{Ma_n}{\varepsilon}\right) \le \frac{\varepsilon}{Ma_n}\mathbb{E}\left\|\boldsymbol{z}_n\right\|^\delta \le \varepsilon$$

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,

$$\begin{aligned}
o_p(1) + o_p(1) &= o_p(1)\\
o_p(1) + O_p(1) &= O_p(1)\\
O_p(1) + O_p(1) &= O_p(1)\\
o_p(1)o_p(1) &= o_p(1)\\
o_p(1)O_p(1) &= o_p(1)\\
O_p(1)O_p(1) &= O_p(1)
\end{aligned}$$

5.12 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 5.12.1 *Assume $\boldsymbol{y}_i$ are independent and identically distributed. If $\mathbb{E}\,|y|^r < \infty$, then as $n \to \infty$*

$$n^{-1/r} \max_{1\leq i\leq n} |y_i| \xrightarrow{p} 0. \tag{5.12}$$

If $\mathbb{E}\exp(ty) < \infty$ for all $t < \infty$, then

$$(\log n)^{-1} \max_{1\leq i\leq n} |y_i| \xrightarrow{p} 0. \tag{5.13}$$

The proof of Theorem 5.12.1 is presented in Section 5.14.

Equivalently, (5.12) can be written as

$$\max_{1\leq i\leq n} |y_i| = o_p(n^{1/r}) \tag{5.14}$$

and (5.13) as

$$\max_{1\leq i\leq n} |y_i| = o_p(\log n). \tag{5.15}$$

Equation (5.12) 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 (5.13) 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 (5.14) as $y_i = o_p(n^{1/r})$ uniformly in $1 \leq i \leq n$, and similarly (5.15) as $y_i = o_p(\log n)$, 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 (5.14)-(5.15).

Theorem 5.12.1 applies to random vectors. If $\mathbb{E}\,\|\boldsymbol{y}\|^r < \infty$ then

$$\max_{1\leq i\leq n} \|\boldsymbol{y}_i\| = o_p(n^{1/r}),$$

and if $\mathbb{E}\exp(\boldsymbol{t}'\boldsymbol{y}) < \infty$ for all $\|\boldsymbol{t}\| < \infty$ then

$$(\log n)^{-1} \max_{1\leq i\leq n} \|\boldsymbol{y}_i\| \xrightarrow{p} 0. \tag{5.16}$$

5.13 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 Section B.11).

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 $\mathbb{E}\|\boldsymbol{y}\|^2 < \infty$ 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 $\mathcal{L}_2 = \left\{F : \mathbb{E}\|\boldsymbol{y}\|^2 < \infty\right\}$ – 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 $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathrm{N}(0,1)$. In this model the maximum likelihood estimator (MLE) $\tilde{\mu}$ for μ is the sample median. Recall from the theory of maximum likelihood that the MLE satisfies $\sqrt{n}(\tilde{\mu} - \mu) \xrightarrow{d} \mathrm{N}\left(0, \left(\mathbb{E}S^2\right)^{-1}\right)$ where $S = \frac{\partial}{\partial \mu}\log f(y \mid \mu) = -\sqrt{2}\,\mathrm{sgn}(y-\mu)$ is the score. We can calculate that $\mathbb{E}S^2 = 2$ and thus conclude that $\sqrt{n}(\tilde{\mu} - \mu) \xrightarrow{d} \mathrm{N}(0, 1/2)$. 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 $\left(\mathbb{E}\boldsymbol{S}_{\boldsymbol{\eta}}\boldsymbol{S}_{\boldsymbol{\eta}}'\right)^{-1}$. Defining $\boldsymbol{M}_\eta = \frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{\mu}_\eta(\boldsymbol{\theta}_0)'$, by Theorem B.11.5 the Cramer-Rao lower bound for estimation of $\boldsymbol{\mu}$ within the submodel η is $\boldsymbol{V}_\eta = \boldsymbol{M}_\eta' \left(\mathbb{E}\boldsymbol{S}_{\boldsymbol{\eta}}\boldsymbol{S}_{\boldsymbol{\eta}}'\right)^{-1} \boldsymbol{M}_\eta$.

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 lower 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 there would exist another submodel η_1 whose Cramer-Rao lower bound satisfies $\boldsymbol{V}_{\eta_0} < \boldsymbol{V}_{\eta_1}$ but this would imply $\boldsymbol{V}_\mu < \boldsymbol{V}_{\eta_1}$ which contradicts the Cramer-Rao Theorem.

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

$$f_\eta(\boldsymbol{y} \mid \boldsymbol{\theta}) = f(\boldsymbol{y})\left(1 + \boldsymbol{\theta}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right) \tag{5.17}$$

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 $\boldsymbol{\theta}$.

[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.

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

$$\boldsymbol{S}_\eta = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_\eta (\boldsymbol{y} \mid \boldsymbol{\theta}_0) = \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) . \tag{5.18}$$

By Theorem B.11.3 the Cramer-Rao lower bound for $\boldsymbol{\theta}$ is

$$\left(\mathbb{E}\left(\boldsymbol{S}_\eta \boldsymbol{S}_\eta'\right)\right)^{-1} = \left(\boldsymbol{V}^{-1} \mathbb{E}\left((\boldsymbol{y} - \boldsymbol{\mu})(\boldsymbol{y} - \boldsymbol{\mu})'\right) \boldsymbol{V}^{-1}\right)^{-1} = \boldsymbol{V}. \tag{5.19}$$

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 (5.17) 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 5.13.1 *In the class of distributions $F \in \mathcal{L}_2$, the semiparametric variance bound for estimation of $\boldsymbol{\mu}$ is $\boldsymbol{V} = \operatorname{var}(y_i)$, 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 5.10.4 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 $\mathcal{L}_2(g) = \left\{F : \mathbb{E}y_1^2 < \infty,\ \mathbb{E}y_2^2 < \infty,\ \text{and } \mathbb{E}y_2 \neq 0\right\}$.

For any submodel η the Cramer-Rao lower bound for estimation of $\boldsymbol{\beta} = \boldsymbol{g}(\boldsymbol{\mu})$ is $\boldsymbol{G}'\boldsymbol{V}_\eta\boldsymbol{G}$ by Theorem B.11.5. For the submodel (5.17) this bound is $\boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}$ which equals the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ from Theorem 5.10.4. Thus $\widehat{\boldsymbol{\beta}}$ is semiparametrically efficient.

Proposition 5.13.2 *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 5.13.2 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.

5.14 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.

Proof of Theorem 5.4.2: 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 $n \geq N$, $\Pr\left(|\overline{y}| > \delta\right) \leq \eta$. Fix δ and η. Set $\varepsilon = \delta\eta/3$. Pick $C < \infty$ large enough so that

$$\mathbb{E}\left(|y_i| \, 1\left(|y_i| > C\right)\right) \leq \varepsilon \tag{5.20}$$

(where $1(\cdot)$ is the indicator function) which is possible since $\mathbb{E}\,|y_i| < \infty$. Define the random variables

$$\begin{aligned} w_i &= y_i 1\left(|y_i| \leq C\right) - \mathbb{E}\left(y_i 1\left(|y_i| \leq C\right)\right) \\ z_i &= y_i 1\left(|y_i| > C\right) - \mathbb{E}\left(y_i 1\left(|y_i| > C\right)\right) \end{aligned}$$

so that

$$\overline{y} = \overline{w} + \overline{z}$$

and

$$\mathbb{E}\,|\overline{y}| \leq \mathbb{E}\,|\overline{w}| + \mathbb{E}\,|\overline{z}|\,. \tag{5.21}$$

We now show that sum of the expectations on the right-hand-side can be bounded below 3ε.

First, by the Triangle Inequality (A.21) and the Expectation Inequality (B.15),

$$\begin{aligned} \mathbb{E}\,|z_i| &= \mathbb{E}\left|y_i 1\left(|y_i| > C\right) - \mathbb{E}\left(y_i 1\left(|y_i| > C\right)\right)\right| \\ &\leq \mathbb{E}\left|y_i 1\left(|y_i| > C\right)\right| + \left|\mathbb{E}\left(y_i 1\left(|y_i| > C\right)\right)\right| \\ &\leq 2\mathbb{E}\left|y_i 1\left(|y_i| > C\right)\right| \\ &\leq 2\varepsilon, \end{aligned} \tag{5.22}$$

and thus by the Triangle Inequality (A.21) and (5.22)

$$\mathbb{E}\,|\overline{z}| = \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} z_i\right| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\,|z_i| \leq 2\varepsilon. \tag{5.23}$$

Second, by a similar argument

$$\begin{aligned} |w_i| &= \left|y_i 1\left(|y_i| \leq C\right) - \mathbb{E}\left(y_i 1\left(|y_i| \leq C\right)\right)\right| \\ &\leq \left|y_i 1\left(|y_i| \leq C\right)\right| + \left|\mathbb{E}\left(y_i 1\left(|y_i| \leq C\right)\right)\right| \\ &\leq 2\left|y_i 1\left(|y_i| \leq C\right)\right| \\ &\leq 2C \end{aligned} \tag{5.24}$$

where the final inequality is (5.20). Then by Jensen's Inequality (B.12), the fact that the w_i are iid and mean zero, and (5.24),

$$\left(\mathbb{E}\,|\overline{w}|\right)^2 \leq \mathbb{E}\,|\overline{w}|^2 = \frac{\mathbb{E}w_i^2}{n} = \frac{4C^2}{n} \leq \varepsilon^2 \tag{5.25}$$

the final inequality holding for $n \geq 4C^2/\varepsilon^2 = 36C^2/\delta^2\eta^2$. Equations (5.21), (5.23) and (5.25) together show that

$$\mathbb{E}\,|\overline{y}| \leq 3\varepsilon^2 \tag{5.26}$$

as desired.

Finally, by Markov's Inequality (B.21) and (5.26),

$$\Pr\left(|\overline{y}| > \delta\right) \leq \frac{\mathbb{E}\,|\overline{y}|}{\delta} \leq \frac{3\varepsilon}{\delta} = \eta,$$

the final equality by the definition of ε. We have shown that for any $\delta > 0$ and $\eta > 0$ then for all $n \geq 36C^2/\delta^2\eta^2$, $\Pr\left(|\overline{y}| > \delta\right) \leq \eta$, as needed. ■

Proof of Theorem 5.6.1: By Loève's c_r Inequality (A.11)

$$\|\boldsymbol{y}\| = \left(\sum_{j=1}^{m} y_j^2\right)^{1/2} \leq \sum_{j=1}^{m} |y_j| \,.$$

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 5.7.1: The moment bound $\mathbb{E}\boldsymbol{y}_i'\boldsymbol{y}_i < \infty$ is sufficient to guarantee that the elements of $\boldsymbol{\mu}$ and $\boldsymbol{V}$ are well defined and finite. Without loss of generality, it is sufficient to consider the case $\boldsymbol{\mu} = \boldsymbol{0}$.

Our proof method is to calculate the characteristic function of $\sqrt{n}\overline{\boldsymbol{y}}_n$ and show that it converges pointwise to the characteristic function of $\mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}\right)$. By Lévy's Continuity Theorem (see Van der Vaart (2008) Theorem 2.13) this is sufficient to established that$\sqrt{n}\overline{\boldsymbol{y}}_n$ converges in distribution to $\mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}\right)$.

For $\boldsymbol{\lambda} \in \mathbb{R}^m$, let $C\left(\boldsymbol{\lambda}\right) = \mathbb{E}\exp\left(\mathrm{i}\boldsymbol{\lambda}'\boldsymbol{y}_i\right)$ denote the characteristic function of $\boldsymbol{y}_i$ and set $c\left(\boldsymbol{\lambda}\right) = \log C(\boldsymbol{\lambda})$. Since $\boldsymbol{y}_i$ has two finite moments the first and second derivatives of $C(\boldsymbol{\lambda})$ are continuous in λ. They are

$$\frac{\partial}{\partial\boldsymbol{\lambda}}C(\boldsymbol{\lambda}) = \mathrm{i}\mathbb{E}\left(\boldsymbol{y}_i\exp\left(i\boldsymbol{\lambda}'\boldsymbol{y}_i\right)\right)$$
$$\frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}C(\boldsymbol{\lambda}) = \mathrm{i}^2\mathbb{E}\left(\boldsymbol{y}_i\boldsymbol{y}_i'\exp\left(i\boldsymbol{\lambda}'\boldsymbol{y}_i\right)\right).$$

When evaluated at $\boldsymbol{\lambda} = \boldsymbol{0}$

$$C(\boldsymbol{0}) = 1$$
$$\frac{\partial}{\partial\boldsymbol{\lambda}}C(\boldsymbol{0}) = i\mathbb{E}\left(\boldsymbol{y}_i\right) = \boldsymbol{0}$$
$$\frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}C(\boldsymbol{0}) = -\mathbb{E}\left(\boldsymbol{y}_i\boldsymbol{y}_i'\right) = -\boldsymbol{V}.$$

Furthermore,

$$c_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}) = \frac{\partial}{\partial\boldsymbol{\lambda}}c(\boldsymbol{\lambda}) = C(\boldsymbol{\lambda})^{-1}\frac{\partial}{\partial\boldsymbol{\lambda}}C(\boldsymbol{\lambda})$$
$$c_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\lambda}) = \frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}c(\boldsymbol{\lambda}) = C(\boldsymbol{\lambda})^{-1}\frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}C(\boldsymbol{\lambda}) - C(\boldsymbol{\lambda})^{-2}\frac{\partial}{\partial\boldsymbol{\lambda}}C\left(\boldsymbol{\lambda}\right)\frac{\partial}{\partial\boldsymbol{\lambda}'}C(\boldsymbol{\lambda})$$

so when evaluated at $\boldsymbol{\lambda} = \boldsymbol{0}$

$$c(\boldsymbol{0}) = 0$$
$$c_{\lambda}(\boldsymbol{0}) = \boldsymbol{0}$$
$$c_{\lambda\lambda}(\boldsymbol{0}) = -\boldsymbol{V}.$$

By a second-order Taylor series expansion of $c(\boldsymbol{\lambda})$ about $\boldsymbol{\lambda} = \mathbf{0}$,

$$c(\boldsymbol{\lambda}) = c(\mathbf{0}) + c_{\lambda}(\mathbf{0})'\boldsymbol{\lambda} + \frac{1}{2}\boldsymbol{\lambda}' c_{\boldsymbol{\lambda\lambda}}(\boldsymbol{\lambda}^*)\boldsymbol{\lambda} = \frac{1}{2}\boldsymbol{\lambda}' c_{\boldsymbol{\lambda\lambda}}(\boldsymbol{\lambda}^*)\boldsymbol{\lambda} \tag{5.27}$$

where $\boldsymbol{\lambda}^*$ lies on the line segment joining $\mathbf{0}$ and $\boldsymbol{\lambda}$.

We now compute $C_n(\boldsymbol{\lambda}) = E\exp\left(\mathrm{i}\boldsymbol{\lambda}'\sqrt{n}\overline{\boldsymbol{y}}_n\right)$, the characteristic function of $\sqrt{n}\overline{\boldsymbol{y}}_n$. By the properties of the exponential function, the independence of the $\boldsymbol{y}_i$, the definition of $c(\boldsymbol{\lambda})$ and (5.27)

$$\begin{aligned}
\log C_n(\boldsymbol{\lambda}) &= \log \mathbb{E}\exp\left(\mathrm{i}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\lambda}'\boldsymbol{y}_i\right) \\
&= \log \mathbb{E}\prod_{i=1}^{n}\exp\left(\mathrm{i}\frac{1}{\sqrt{n}}\boldsymbol{\lambda}'\boldsymbol{y}_i\right) \\
&= \log \prod_{i=1}^{n}\mathbb{E}\exp\left(\mathrm{i}\frac{1}{\sqrt{n}}\boldsymbol{\lambda}'\boldsymbol{y}_i\right) \\
&= \sum_{i=1}^{n}\log \mathbb{E}\exp\left(\mathrm{i}\frac{1}{\sqrt{n}}\boldsymbol{\lambda}'\boldsymbol{y}_i\right) \\
&= nc\left(\frac{\boldsymbol{\lambda}}{\sqrt{n}}\right) \\
&= \frac{1}{2}\boldsymbol{\lambda}' c_{\boldsymbol{\lambda\lambda}}(\boldsymbol{\lambda}_n)\boldsymbol{\lambda}
\end{aligned}$$

where $\boldsymbol{\lambda}_n$ lies on the line segment joining $\mathbf{0}$ and $\boldsymbol{\lambda}/\sqrt{n}$. Since $\boldsymbol{\lambda}_n \to 0$ and $c_{\boldsymbol{\lambda\lambda}}(\boldsymbol{\lambda})$ is continuous, $c_{\boldsymbol{\lambda\lambda}}(\boldsymbol{\lambda}_n) \to c_{\boldsymbol{\lambda\lambda}}(\mathbf{0}) = -\boldsymbol{V}$. We thus find that as $n \to \infty$,

$$\log C_n(\boldsymbol{\lambda}) \to -\frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{V}\boldsymbol{\lambda}$$

and

$$C_n(\boldsymbol{\lambda}) \to \exp\left(-\frac{1}{2}\boldsymbol{\lambda}'\boldsymbol{V}\boldsymbol{\lambda}\right)$$

which is the characteristic function of the $\mathrm{N}(\mathbf{0}, \boldsymbol{V})$ distribution. This completes the proof. ■

Proof of Theorem 5.9.1: Since $\boldsymbol{g}$ is continuous at $\boldsymbol{c}$, for all $\varepsilon > 0$ we can find a $\delta > 0$ such that if $\|\boldsymbol{z}_n - \boldsymbol{c}\| < \delta$ then $\|\boldsymbol{g}(\boldsymbol{z}_n) - \boldsymbol{g}(\boldsymbol{c})\| \le \varepsilon$. Recall that $A \subset B$ implies $\Pr(A) \le \Pr(B)$. Thus $\Pr(\|\boldsymbol{g}(\boldsymbol{z}_n) - \boldsymbol{g}(\boldsymbol{c})\| \le \varepsilon) \ge \Pr(\|\boldsymbol{z}_n - \boldsymbol{c}\| < \delta) \to 1$ as $n \to \infty$ by the assumption that $\boldsymbol{z}_n \xrightarrow{p} \boldsymbol{c}$. Hence $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{p} \boldsymbol{g}(\boldsymbol{c})$ as $n \to \infty$. ■

Proof of Theorem 5.10.3: By a vector Taylor series expansion, for each element of $\boldsymbol{g}$,

$$g_j(\boldsymbol{\theta}_n) = g_j(\boldsymbol{\theta}) + g_{j\boldsymbol{\theta}}(\boldsymbol{\theta}^*_{jn})(\boldsymbol{\theta}_n - \boldsymbol{\theta})$$

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

$$\sqrt{n}\left(\boldsymbol{g}(\boldsymbol{\theta}_n) - \boldsymbol{g}(\boldsymbol{\theta})\right) = (\boldsymbol{G} + a_n)'\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{G}'\boldsymbol{\xi}. \tag{5.28}$$

The convergence is by Theorem 5.10.1, as $\boldsymbol{G} + a_n \xrightarrow{d} \boldsymbol{G}$, $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{\xi}$, and their product is continuous. This establishes (5.10)

When $\boldsymbol{\xi} \sim \mathrm{N}(0, \boldsymbol{V})$, the right-hand-side of (5.28) equals

$$\boldsymbol{G}'\xi = \boldsymbol{G}'\,\mathrm{N}(0, \boldsymbol{V}) = \mathrm{N}\left(0, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}\right)$$

establishing (5.11). ■

Proof of Theorem 5.12.1: First consider (5.12). Take any $\delta > 0$. The event $\left\{\max_{1\le i\le n} |y_i| > \delta n^{1/r}\right\}$ 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,

$$\begin{aligned}
\Pr\left(n^{-1/r} \max_{1\le i\le n} |y_i| > \delta\right) &= \Pr\left(\bigcup_{i=1}^n \{|y_i|^r > \delta^r n\}\right) \\
&\le \sum_{i=1}^n \Pr\left(|y_i|^r > n\delta^r\right) \\
&\le \frac{1}{n\delta^r} \sum_{i=1}^n \mathbb{E}\left(|y_i|^r 1\left(|y_i|^r > n\delta^r\right)\right) \\
&= \frac{1}{\delta^r} \mathbb{E}\left(|y_i|^r 1\left(|y_i|^r > n\delta^r\right)\right)
\end{aligned}$$

where the second inequality is the strong form of Markov's inequality (Theorem B.22) and the final equality is since the y_i are iid. Since $\mathbb{E}|y|^r < \infty$ this final expectation converges to zero as $n \to \infty$. This is because

$$\mathbb{E}|y_i|^r = \int |y|^r \, dF(y) < \infty$$

implies

$$\mathbb{E}\left(|y_i|^r 1\left(|y_i|^r > c\right)\right) = \int_{|y|^r > c} |y|^r \, dF(y) \to 0 \tag{5.29}$$

as $c \to \infty$. This establishes (5.12).

Now consider (5.13). Take any $\delta > 0$ and set $t = 1/\delta$. By a similar calculation

$$\begin{aligned}
\Pr\left((\log n)^{-1} \max_{1\le i\le n} |y_i| > \delta\right) &= \Pr\left(\bigcup_{i=1}^n \{\exp|ty_i| > \exp(t\delta \log n)\}\right) \\
&\le \sum_{i=1}^n \Pr\left(\exp|ty_i| > n\right) \\
&\le \mathbb{E}\left(\exp|ty| \, 1\left(\exp|ty| > n\right)\right)
\end{aligned}$$

where the second line uses $\exp(t\delta \log n) = \exp(\log n) = n$. 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 (5.29). This establishes (5.13). ■

Exercises

Exercise 5.1 For the following sequences, find the liminf, limsup and limit (if it exists) as $n \to \infty$

1. $a_n = 1/n$
2. $a_n = \sin\left(\frac{\pi}{2}n\right)$
3. $a_n = \frac{1}{n}\sin\left(\frac{\pi}{2}n\right)$

Exercise 5.2 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.

1. Show that $\overline{y}^*$ is unbiased for $\mu = \mathbb{E}y_i$.
2. Calculate $\text{var}(\overline{y}^*)$.
3. Show that a sufficient condition for $\overline{y}^* \xrightarrow{p} \mu$ is that $\frac{1}{n^2}\sum_{i=1}^n w_i^2 \longrightarrow 0$.
4. Show that a sufficient condition for the condition in part 3 is $\max_{i\leq n} w_i = o(n)$.

Exercise 5.3 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 5.4 Find the moment estimator $\widehat{\mu}_3$ of $\mu_3 = \mathbb{E}y_i^3$ and show that $\sqrt{n}\,(\widehat{\mu}_3 - \mu_3) \xrightarrow{d} \text{N}\left(0, v^2\right)$ for some v^2. Write v^2 as a function of the moments of y_i.

Exercise 5.5 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 5.6 Suppose $\sqrt{n}\,(\widehat{\mu} - \mu) \xrightarrow{d} \text{N}\left(0, v^2\right)$ and set $\beta = \mu^2$ and $\widehat{\beta} = \widehat{\mu}^2$.

1. Use the Delta Method to obtain an asymptotic distribution for $\sqrt{n}\left(\widehat{\beta} - \beta\right)$.
2. Now suppose $\mu = 0$. Describe what happens to the asymptotic distribution from the previous part.
3. Improve on the previous answer. Under the assumption $\mu = 0$, find the asymptotic distribution for $n\widehat{\beta} = n\widehat{\mu}^2$.
4. Comment on the differences between the answers in parts 1 and 3.

Chapter 6

Asymptotic Theory for Least Squares

6.1 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 1.5.1) and finite second moments (Assumption 2.18.1). We restate this condition here for clarity.

Assumption 6.1.1

1. *The observations* $(y_i, \boldsymbol{x_i})$, $i = 1, ..., n$, *are independent and identically distributed.*
2. $\mathbb{E}y^2 < \infty$.
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 6.1.2 *In addition to Assumption 6.1.1,* $\mathbb{E}y_i^4 < \infty$ *and* $\mathbb{E}\left\|\boldsymbol{x}_i\right\|^4 < \infty$.

6.2 Consistency of Least-Squares Estimator

In this section we use the weak law of large numbers (WLLN, Theorem 5.4.2 and Theorem 5.6.2) and continuous mapping theorem (CMT, Theorem 5.9.1) 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

$$\widehat{\boldsymbol{\beta}}=\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_iy_i\right)=\widehat{\boldsymbol{Q}}_{xx}^{-1}\widehat{\boldsymbol{Q}}_{xy}$$

is a function of the sample moments $\widehat{\boldsymbol{Q}}_{xx}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'$ and $\widehat{\boldsymbol{Q}}_{xy}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_iy_i$.

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 $\boldsymbol{x}_iy_i$. These variables also have finite expectations by Theorem 2.18.1.1. Under these conditions, the WLLN (Theorem 5.6.2) implies that as $n\to\infty$,

$$\widehat{\boldsymbol{Q}}_{xx}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\xrightarrow{p}\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)=\boldsymbol{Q}_{xx} \tag{6.1}$$

and

$$\widehat{\boldsymbol{Q}}_{xy}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_iy_i\xrightarrow{p}\mathbb{E}\left(\boldsymbol{x}_iy_i\right)=\boldsymbol{Q}_{xy}. \tag{6.2}$$

Third, the CMT (Theorem 5.9.1) allows us to combine these equations to show that $\widehat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$. Specifically, as $n\to\infty$,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}&=\widehat{\boldsymbol{Q}}_{xx}^{-1}\widehat{\boldsymbol{Q}}_{xy}\\&\xrightarrow{p}\boldsymbol{Q}_{xx}^{-1}\boldsymbol{Q}_{xy}\\&=\boldsymbol{\beta}.\end{aligned} \tag{6.3}$$

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

$$\widehat{\boldsymbol{\beta}}=\boldsymbol{g}\left(\widehat{\boldsymbol{Q}}_{xx},\widehat{\boldsymbol{Q}}_{xy}\right)$$

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 2.18.1 implies that $\boldsymbol{Q}_{xx}^{-1}$ exists and thus $\boldsymbol{g}(\boldsymbol{A},\boldsymbol{b})$ is continuous at $\boldsymbol{A}=\boldsymbol{Q}_{xx}$. This justifies the application of the CMT in (6.3).

For a slightly different demonstration of (6.3), recall that (4.7) implies that

$$\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\widehat{\boldsymbol{Q}}_{xx}^{-1}\widehat{\boldsymbol{Q}}_{xe} \tag{6.4}$$

where

$$\widehat{\boldsymbol{Q}}_{xe}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_ie_i.$$

The WLLN and (2.27) imply

$$\widehat{\boldsymbol{Q}}_{xe} \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i e_i\right) = \boldsymbol{0}. \tag{6.5}$$

Therefore

$$\begin{aligned}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \widehat{\boldsymbol{Q}}_{xx}^{-1}\widehat{\boldsymbol{Q}}_{xe} \\ &\xrightarrow{p} \boldsymbol{Q}_{xx}^{-1}\boldsymbol{0} \\ &= \boldsymbol{0}\end{aligned}$$

which is the same as $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

Theorem 6.2.1 ***Consistency of Least-Squares***
Under Assumption 6.1.1, $\widehat{\boldsymbol{Q}}_{xx} \xrightarrow{p} \boldsymbol{Q}_{xx}$, $\widehat{\boldsymbol{Q}}_{xy} \xrightarrow{p} \boldsymbol{Q}_{xy}$, $\widehat{\boldsymbol{Q}}_{xx}^{-1} \xrightarrow{p} \boldsymbol{Q}_{xx}^{-1}$, $\widehat{\boldsymbol{Q}}_{xe} \xrightarrow{p} \boldsymbol{0}$, *and* $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ *as* $n \to \infty$.

Theorem 6.2.1 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 6.2.1 can be equivalently written as

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + o_p(1). \tag{6.6}$$

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 6.1. 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 $\hat{\beta}_1 = 0.114$.

6.3 Asymptotic Normality

We started this chapter discussing the need for an approximation to the distribution of the OLS estimator $\widehat{\boldsymbol{\beta}}$. In Section 6.2 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 (6.4) and multiply it by $\sqrt{n}$. This yields the expression

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{x}_i e_i\right). \tag{6.7}$$

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 5.7.1) applies.

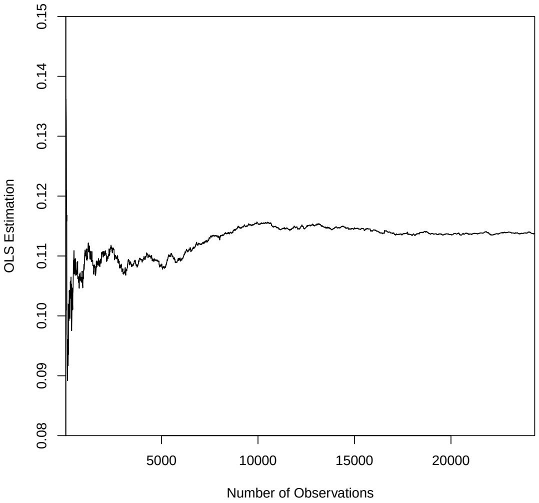


Figure 6.1: The least-squares estimator $\hat{\beta}_1$ as a function of sample size n

The product $\boldsymbol{x}_i e_i$ is iid (since the observations are iid) and mean zero (since $\mathbb{E}(\boldsymbol{x}_i e_i) = \boldsymbol{0}$). Define the $k \times k$ covariance matrix

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right). \tag{6.8}$$

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 6.1.2 implies that $\mathbb{E}e_i^4 < \infty$.

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.15), the $j\ell^{th}$ element of $\boldsymbol{\Omega}$ is

$$\left|\mathbb{E}\left(x_{ji} x_{\ell i} e_i^2\right)\right| \leq \mathbb{E}\left|x_{ji} x_{\ell i} e_i^2\right| = \mathbb{E}\left(|x_{ji}| \, |x_{\ell i}| \, e_i^2\right).$$

By two applications o fthe Cauchy-Schwarz Inequality (B.17), this is smaller than

$$\left(\mathbb{E}\left(x_{ji}^2 x_{\ell i}^2\right)\right)^{1/2} \left(\mathbb{E}e_i^4\right)^{1/2} \leq \left(\mathbb{E}x_{ji}^4\right)^{1/4} \left(\mathbb{E}x_{\ell i}^4\right)^{1/4} \left(\mathbb{E}e_i^4\right)^{1/2} < \infty$$

where the finiteness holds under Assumption 6.1.2.

An alternative way to show that the elements of $\boldsymbol{\Omega}$ are finite is by using a matrix norm $\|\cdot\|$ (See Appendix A.13). Then by the Expectation Inequality, the Cauchy-Schwarz Inequality, and Assumption 6.1.2

$$\|\boldsymbol{\Omega}\| \leq \mathbb{E}\left\|\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right\| = \mathbb{E}\left(\|\boldsymbol{x}_i\|^2 e_i^2\right) \leq \left(\mathbb{E}\|\boldsymbol{x}_i\|^4\right)^{1/2} \left(\mathbb{E}e_i^4\right)^{1/2} < \infty.$$

This is a more compact argument (often described as more *elegant*) but it 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 5.7.1).

Theorem 6.3.1 *Under Assumption 6.1.2,*

$$\boldsymbol{\Omega} < \infty \tag{6.9}$$

and

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{x}_i e_i \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}) \tag{6.10}$$

as $n \to \infty$.

Putting together (6.1), (6.7), and (6.10),

$$\begin{aligned} \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) &\xrightarrow{d} \boldsymbol{Q}_{xx}^{-1} \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega}) \\ &= \mathrm{N}\left(\mathbf{0}, \boldsymbol{Q}_{xx}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}_{xx}^{-1}\right) \end{aligned}$$

as $n \to \infty$, where the final equality follows from the property that linear combinations of normal vectors are also normal (Theorem B.9.1).

We have derived the asymptotic normal approximation to the distribution of the least-squares estimator.

Theorem 6.3.2 ***Asymptotic Normality of Least-Squares Estimator***

Under Assumption 6.1.2, as $n \to \infty$

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}})$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{xx}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}_{xx}^{-1}, \tag{6.11}$$

$\boldsymbol{Q}_{xx} = \mathbb{E}(\boldsymbol{x}_i \boldsymbol{x}_i')$, *and* $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i \boldsymbol{x}_i' e_i^2\right)$.

In the stochastic order notation, Theorem 6.3.2 implies that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + O_p(n^{-1/2}) \tag{6.12}$$

which is stronger than (6.6).

The matrix $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{xx}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}_{xx}^{-1}$ 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 $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{xx}^{-1} \boldsymbol{\Omega} \boldsymbol{Q}_{xx}^{-1}$ is called a **sandwich** form, as the matrix $\boldsymbol{\Omega}$ is sandwiched between two copies of $\boldsymbol{Q}_{xx}^{-1}$.

It is useful to compare the variance of the asymptotic distribution given in (6.11) and the finite-sample conditional variance in the CEF model as given in (4.12):

$$\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} = \mathrm{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right) \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{6.13}$$

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 (6.13) 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. We say that e_i is a **Homoskedastic Projection Error** when

$$\operatorname{cov}(\boldsymbol{x}_i\boldsymbol{x}_i', e_i^2) = \boldsymbol{0}. \tag{6.14}$$

Condition (6.14) holds in the homoskedastic linear regression model, but is somewhat broader. Under (6.14) the asymptotic variance formulae simplify as

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\mathbb{E}\left(e_i^2\right) = \boldsymbol{Q}_{xx}\sigma^2 \tag{6.15}$$

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{xx}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{xx}^{-1} = \boldsymbol{Q}_{xx}^{-1}\sigma^2 \equiv \boldsymbol{V}_{\boldsymbol{\beta}}^0 \tag{6.16}$$

In (6.16) we define $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{xx}^{-1}\sigma^2$ whether (6.14) is true or false. When (6.14) 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 6.3.2 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 6.1.2, 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 5.1 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 $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\hat{\beta}_1 - \beta_1\right)$ has the $\mathrm{N}(0,1)$ asymptotic distibution for any $\alpha > 2$. In Figure 6.2 we display the finite sample densities of the normalized estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\hat{\beta}_1 - \beta_1\right)$, 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 6.3. Here the model is $y_i = \beta + e_i$ where

$$e_i = \frac{u_i^k - \mathbb{E}\left(u_i^k\right)}{\left(\mathbb{E}\left(u_i^{2k}\right) - \left(\mathbb{E}\left(u_i^k\right)\right)^2\right)^{1/2}} \tag{6.17}$$

and $u_i \sim \mathrm{N}(0,1)$. We show the sampling distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ setting $n = 100$, for $k = 1$, 4, 6 and 8. As k increases, the sampling distribution becomes highly skewed and non-normal. The lesson from Figures 6.2 and 6.3 is that the $\mathrm{N}(0,1)$ asymptotic approximation is never guaranteed to be accurate.

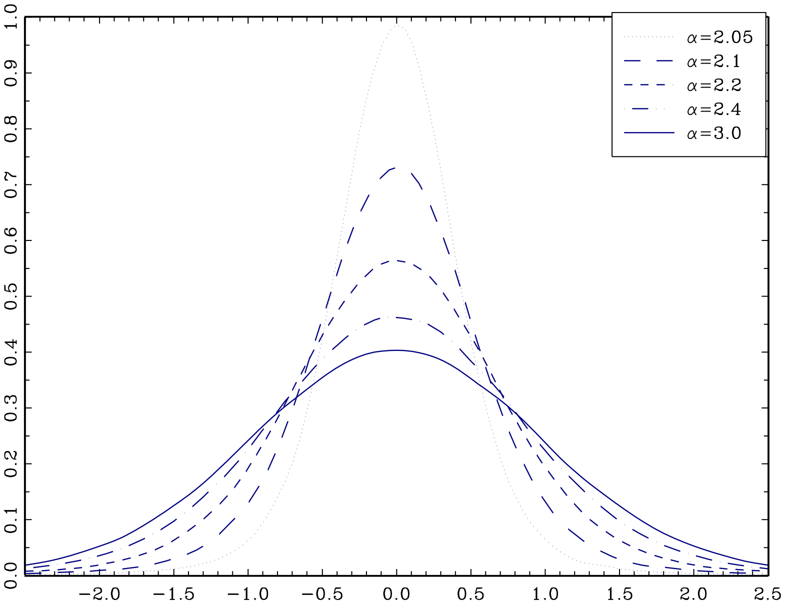


Figure 6.2: Density of Normalized OLS estimator with Double Pareto Error

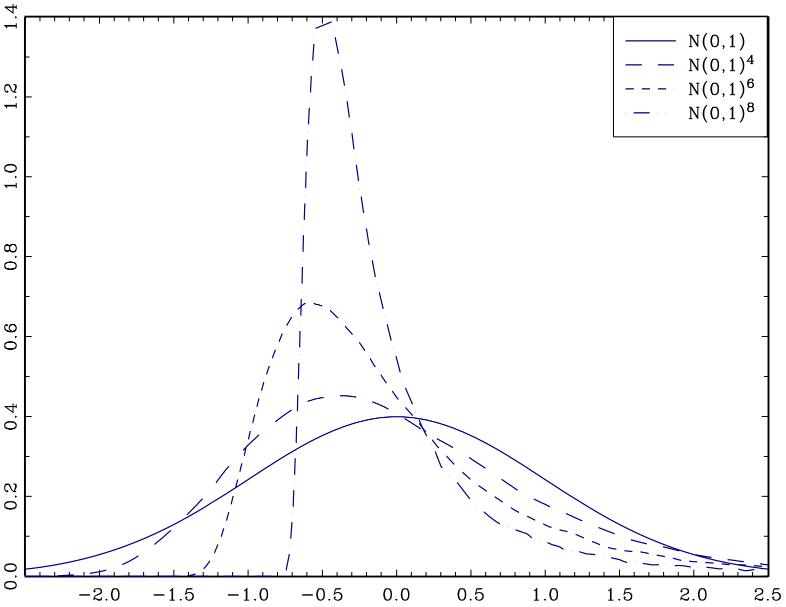


Figure 6.3: Density of Normalized OLS estimator with error process (6.17)

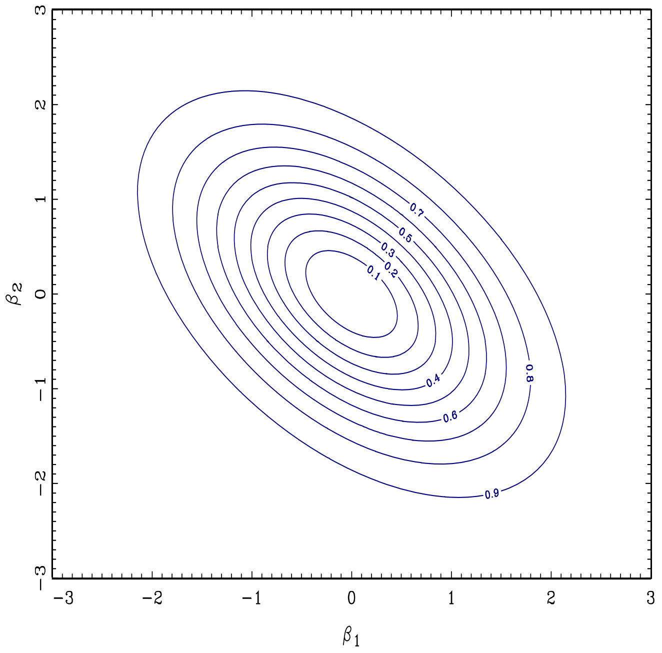


Figure 6.4: Contours of Joint Distribution of $(\hat\beta_1, \hat\beta_2)$, homoskedastic case

6.4 Joint Distribution

Theorem 6.3.2 gives the joint asymptotic distribution of the coefficient estimates. We can use the result to study the covariance between the coefficient estimates. For simplicity, supose $k = 2$ with no intercept, both regressors are mean zero and the error is homoskedastic. Let σ_1^2 and σ_2^2 be the variances of x_{1i} and x_{2i}, and ρ be their correlation. Then using the formula for inversion of a 2×2 matrix,

$$\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \sigma^2 \boldsymbol{Q}_{xx}^{-1} = \frac{\sigma^2}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)} \left[\begin{array}{cc} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{array} \right].$$

Thus if x_{1i} and x_{2i} are positively correlated ($\rho > 0$) then $\hat\beta_1$ and $\hat\beta_2$ are negatively correlated (and vice-versa).

For illustration, Figure 6.4 displays the probability contours of the joint asymptotic distribution of $\hat\beta_1 - \beta_1$ and $\hat\beta_2 - \beta_2$ 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 $\hat\beta_1$ is unusually negative, it is likely that $\hat\beta_2$ is unusually positive, or conversely. It is also unlikely that we will observe both $\hat\beta_1$ and $\hat\beta_2$ 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 $\Pr\left(x_{1i} = x_{2i} = 1\right) = \Pr\left(x_{1i} = x_{2i} = -1\right) = 3/8$, and $\Pr\left(x_{1i} = 1, x_{2i} = -1\right) = \Pr\left(x_{1i} = -1, x_{2i} = 1\right) = 1/8$. You can check that the regressors are mean zero, unit variance and correlation 0.5, which is identical with the setting displayed in Figure 6.4.

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

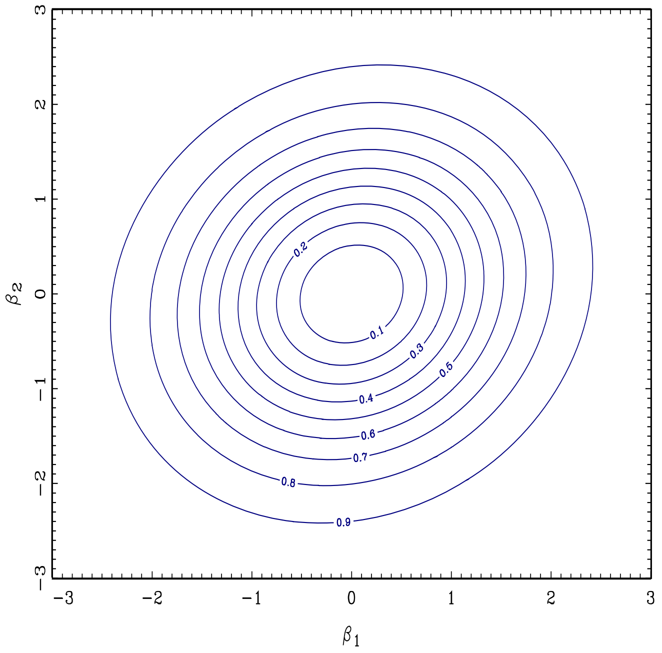


Figure 6.5: Contours of Joint Distribution of $\hat{\beta}_1$ and $\hat{\beta}_2$, heteroskedastic case

$\mathbb{E}\left(x_{1i}x_{2i}e_i^2\right) = \dfrac{7}{8}$. Therefore

$$
\begin{aligned}
\boldsymbol{V}_{\boldsymbol{\beta}} &= \boldsymbol{Q}_{xx}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{xx}^{-1} \\
&= \frac{9}{16}\left[\begin{array}{cc} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{array}\right]\left[\begin{array}{cc} 1 & \frac{7}{8} \\ \frac{7}{8} & 1 \end{array}\right]\left[\begin{array}{cc} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{array}\right] \\
&= \frac{4}{3}\left[\begin{array}{cc} 1 & \frac{1}{4} \\ \frac{1}{4} & 1 \end{array}\right].
\end{aligned}
$$

Thus the coefficient estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are positively correlated (their correlation is 1/4.) The joint probability contours of their asymptotic distribution is displayed in Figure 6.5. 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

$$
\boldsymbol{Q}_{xx} = \left[\begin{array}{cc} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{array}\right], \qquad \boldsymbol{\Omega} = \left[\begin{array}{cc} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{array}\right]. \tag{6.18}
$$

From (2.41)

$$
\boldsymbol{Q}_{xx}^{-1} = \left[\begin{array}{cc} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{array}\right]
$$

where $\boldsymbol{Q}_{11\cdot 2} = \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and $\boldsymbol{Q}_{22\cdot 1} = \boldsymbol{Q}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$. Thus when the error is homoskedastic,

$$\operatorname{cov}\left(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2\right) = -\sigma^2 \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}$$

which is a matrix generalization of the two-regressor case.

In the general case, you can show that (Exercise 6.5)

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix} \tag{6.19}$$

where

$$\boldsymbol{V}_{11} = \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\boldsymbol{\Omega}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{\Omega}_{21} - \boldsymbol{\Omega}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{\Omega}_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot 2}^{-1} \tag{6.20}$$

$$\boldsymbol{V}_{21} = \boldsymbol{Q}_{22\cdot 1}^{-1}\left(\boldsymbol{\Omega}_{21} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot 2}^{-1} \tag{6.21}$$

$$\boldsymbol{V}_{22} = \boldsymbol{Q}_{22\cdot 1}^{-1}\left(\boldsymbol{\Omega}_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{12} - \boldsymbol{\Omega}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12} + \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{11}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\right)\boldsymbol{Q}_{22\cdot 1}^{-1} \tag{6.22}$$

Unfortunately, these expressions are not easily interpretable.

6.5 Consistency of Error Variance Estimators

Using the methods of Section 6.2 we can show that the estimators $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2$ and $s^2 = \frac{1}{n-k}\sum_{i=1}^n \hat{e}_i^2$ are consistent for σ^2.

The trick is to write the residual $\hat{e}_i$ as equal to the error e_i plus a deviation term

$$\begin{aligned} \hat{e}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \\ &= e_i + \boldsymbol{x}_i'\boldsymbol{\beta} - x_i'\widehat{\boldsymbol{\beta}} \\ &= e_i - \boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \end{aligned}$$

Thus the squared residual equals the squared error plus a deviation

$$\hat{e}_i^2 = e_i^2 - 2e_i\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{x}_i\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{6.23}$$

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.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^n e_i\boldsymbol{x}_i'\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{6.24}$$

Indeed, the WLLN shows that

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n e_i^2 \xrightarrow{p} \sigma^2 \\ &\frac{1}{n}\sum_{i=1}^n e_i\boldsymbol{x}_i' \xrightarrow{p} \mathbb{E}\left(e_i\boldsymbol{x}_i'\right) = 0 \\ &\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' \xrightarrow{p} \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) = \boldsymbol{Q}_{xx} \end{aligned}$$

and Theorem 6.2.1 shows that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$. Hence (6.24) converges in probability to σ^2, as desired.

Finally, since $n/(n-k) \to 1$ as $n \to \infty$, it follows that

$$s^2 = \left(\frac{n}{n-k}\right)\hat{\sigma}^2 \xrightarrow{p} \sigma^2.$$

Thus both estimators are consistent.

Theorem 6.5.1 *Under Assumption 6.1.1,* $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ *and* $s^2 \xrightarrow{p} \sigma^2$ *as* $n \to \infty$.

6.6 Homoskedastic Covariance Matrix Estimation

Theorem 6.3.2 shows that $\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$ is asymptotically normal with 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 (6.1), 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 (4.21). 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 6.2.1 established that $\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}} \xrightarrow{p} \boldsymbol{Q}_{\boldsymbol{xx}}$, and Theorem 6.5.1 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 6.1.1.4. 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 6.6.1 *Under Assumption 6.1.1,* $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0 \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}^0$ *as* $n \to \infty$.

It is instructive to notice that Theorem 6.6.1 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.

6.7 Heteroskedastic Covariance Matrix Estimation

Theorems 6.3.2 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 (6.1).

The moment estimator for $\boldsymbol{\Omega}$ is

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2, \tag{6.25}$$

leading to the plug-in covariance matrix estimator

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^W = \widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\widehat{\boldsymbol{\Omega}}\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}. \tag{6.26}$$

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 (4.28).

As shown in Theorem 6.2.1, $\widehat{\boldsymbol{Q}}_{xx}^{-1} \xrightarrow{p} \boldsymbol{Q}_{xx}^{-1}$, so we just need to verify the consistency of $\widehat{\boldsymbol{\Omega}}$. The key is to replace the squared residual $\hat{e}_i^2$ with the squared error e_i^2, and then show that the difference is asymptotically negligible.

Specifically, observe that

$$\begin{aligned}\widehat{\boldsymbol{\Omega}} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left(\hat{e}_i^2 - e_i^2\right). \qquad (6.27)\end{aligned}$$

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 (6.9).

So to establish that $\widehat{\boldsymbol{\Omega}}$ is consistent for $\boldsymbol{\Omega}$ it remains to show that

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left(\hat{e}_i^2 - e_i^2\right) \xrightarrow{p} 0. \qquad (6.28)$$

There are multiple ways to do this. A reasonable straightforward yet slightly tedious derivation is to start by applying the Triangle Inequality (A.21) using a matrix norm:

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left(\hat{e}_i^2 - e_i^2\right) \right\| &\leq \frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{x}_i \boldsymbol{x}_i' \left(\hat{e}_i^2 - e_i^2\right) \right\| \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\| \boldsymbol{x}_i \right\|^2 \left| \hat{e}_i^2 - e_i^2 \right|. \qquad (6.29)\end{aligned}$$

Then recalling the expression for the squared residual (6.23), apply the Triangle Inequality and then the Schwarz Inequality (A.15) twice

$$\begin{aligned}\left|\hat{e}_i^2 - e_i^2\right| &\leq 2\left| e_i \boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right| + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{x}_i \boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\ &= 2\left|e_i\right| \left|\boldsymbol{x}_i' \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right| + \left|\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{x}_i\right|^2 \\ &\leq 2\left|e_i\right| \left\|\boldsymbol{x}_i\right\| \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| + \left\|\boldsymbol{x}_i\right\|^2 \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2. \qquad (6.30)\end{aligned}$$

Combining (6.29) and (6.30), we find

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i' \left(\hat{e}_i^2 - e_i^2\right) \right\| &\leq 2\left(\frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i\right\|^3 \left|e_i\right|\right) \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| + \left(\frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i\right\|^4\right) \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2 \\ &= o_p(1). \qquad (6.31)\end{aligned}$$

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 6.1.2 (and are thus $O_p(1)$). Indeed, by Hölder's Inequality (B.16)

$$\mathbb{E}\left(\left\|\boldsymbol{x}_i\right\|^3 \left|e_i\right|\right) \leq \left(\mathbb{E}\left(\left\|\boldsymbol{x}_i\right\|^3\right)^{4/3}\right)^{3/4} \left(\mathbb{E}e_i^4\right)^{1/4} = \left(\mathbb{E}\left\|\boldsymbol{x}_i\right\|^4\right)^{3/4} \left(\mathbb{E}e_i^4\right)^{1/4} < \infty.$$

We have established (6.28), as desired.

Theorem 6.7.1 *Under Assumption 6.1.2, as* $n \to \infty$, $\widehat{\mathbf{\Omega}} \xrightarrow{p} \mathbf{\Omega}$ *and* $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$.

For an alternative proof of this result, see Section 6.21.

6.8 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

$$\begin{aligned}\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}} &= \operatorname{var}\left(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ \boldsymbol{V}_{\boldsymbol{\beta}} &= \operatorname{avar}\left(\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right) = \boldsymbol{Q}_{xx}^{-1}\mathbf{\Omega}\boldsymbol{Q}_{xx}^{-1}.\end{aligned}$$

The White estimates of these two covariance matrices are

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{W} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{W} &= \widehat{\boldsymbol{Q}}_{xx}^{-1}\widehat{\mathbf{\Omega}}\widehat{\boldsymbol{Q}}_{xx}^{-1}\end{aligned}$$

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

$$\begin{aligned}\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}^{0} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\ \boldsymbol{V}_{\boldsymbol{\beta}}^{0} &= \boldsymbol{Q}_{xx}^{-1}\sigma^2\end{aligned}$$

and their standard estimators are

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}^{0} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}s^2 \\ \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{0} &= \widehat{\boldsymbol{Q}}_{xx}^{-1}s^2\end{aligned}$$

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.

6.9 Alternative Covariance Matrix Estimators*

In Section 6.7 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 4.11, 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 (6.26) and shown consistent for $\boldsymbol{V}_{\boldsymbol{\beta}}$ in Theorem 6.7.1. 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 (6.26) but with $\widehat{\boldsymbol{\Omega}}$ replaced by

$$\widetilde{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-2}\,\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2$$

and

$$\overline{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-1}\,\boldsymbol{x}_i\boldsymbol{x}_i'\hat{e}_i^2,$$

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:

$$h_n^* = \max_{1\le i\le n} h_{ii} = o_p(1). \tag{6.32}$$

(See Theorem 6.22.1 in Section 6.22).) Then using the Triangle Inequality

$$\begin{aligned}\left\|\overline{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}}\right\| &\le \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\boldsymbol{x}_i'\right\|\hat{e}_i^2\left|(1-h_{ii})^{-1}-1\right| \\ &\le \left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|^2\hat{e}_i^2\right)\left|(1-h_n^*)^{-1}-1\right|.\end{aligned}$$

The sum in parenthesis can be shown to be $O_p(1)$ under Assumption 6.1.2 by the same argument as in in the proof of Theorem 6.7.1. (In fact, it can be shown to converge in probability to $\mathbb{E}\left(\left\|\boldsymbol{x}_i\right\|^2 e_i^2\right)$.) The term in absolute values is $o_p(1)$ by (6.32). Thus the product is $o_p(1)$, which means that $\overline{\boldsymbol{\Omega}} = \widehat{\boldsymbol{\Omega}} + o_p(1) \longrightarrow \boldsymbol{\Omega}$.

Similarly,

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\Omega}} - \widehat{\boldsymbol{\Omega}}\right\| &\le \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\boldsymbol{x}_i'\right\|\hat{e}_i^2\left|(1-h_{ii})^{-2}-1\right| \\ &\le \left(\frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i\right\|^2\hat{e}_i^2\right)\left|(1-h_n^*)^{-2}-1\right| \\ &= o_p(1).\end{aligned}$$

Theorem 6.9.1 *Under Assumption 6.1.2, 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 6.9.1 shows that the alternative covariance matrix estimators are also consistent for the asymptotic covariance matrix.

6.10 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, we 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 5.9.1) 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 6.10.1 *Under Assumption 6.1.1, 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 5.10.3) we can show that $\widehat{\boldsymbol{\theta}}$ is asymptotically normal.

Assumption 6.10.1 *$\boldsymbol{r}(\boldsymbol{\beta}) : \mathbb{R}^k \to \mathbb{R}^q$ 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 6.10.2 ***Asymptotic Distribution of Functions of Parameters***
Under Assumptions 6.1.2 and 6.10.1, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right) \tag{6.33}$$

where

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \boldsymbol{R}' \boldsymbol{V}_{\beta} \boldsymbol{R} \tag{6.34}$$

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"

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{pmatrix} \tag{6.35}$$

then we can conformably 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

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \end{pmatrix} \boldsymbol{V}_{\boldsymbol{\beta}} \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{pmatrix} = \boldsymbol{V}_{11},$$

the upper-left sub-matrix of $\boldsymbol{V}_{11}$ given in (6.20). In this case (6.33) states that

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{11}\right).$$

That is, subsets of $\widehat{\boldsymbol{\beta}}$ are approximately normal with variances given by the comformable 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

$$\boldsymbol{R} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial}{\partial \beta_1}(\beta_j/\beta_l) \\ \vdots \\ \frac{\partial}{\partial \beta_j}(\beta_j/\beta_l) \\ \vdots \\ \frac{\partial}{\partial \beta_\ell}(\beta_j/\beta_l) \\ \vdots \\ \frac{\partial}{\partial \beta_k}(\beta_j/\beta_l) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1/\beta_l \\ \vdots \\ -\beta_j/\beta_l^2 \\ \vdots \\ 0 \end{pmatrix} \tag{6.36}$$

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

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}(\widehat{\boldsymbol{\beta}})'. \tag{6.37}$$

The derivative in (6.37) 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 (6.36). 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

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{R}}. \tag{6.38}$$

Alternatively, $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^0$, $\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}$ 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

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^0 = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{\boldsymbol{0}}\widehat{\boldsymbol{R}} = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{Q}}_{xx}^{-1}\widehat{\boldsymbol{R}}s^2 \tag{6.39}$$

Given (6.37), (6.38) and (6.39) are simple to calculate using matrix operations.

As the primary justification for $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ is the asymptotic approximation (6.33), $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ 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 6.10.2 since $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\beta}}$ by Theorem 6.7.1, 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.

Theorem 6.10.3 *Under Assumptions 6.1.2 and 6.10.1, as $n \to \infty$,*

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\theta}}.$$

Theorem 6.10.3 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

$$\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}} = n^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}} \tag{6.40}$$

as an estimate of the variance of $\widehat{\boldsymbol{\theta}}$, or substitute an alternative covariance estimator such as $\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$.

6.11 Asymptotic Standard Errors

As described in Section 4.12, 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

$$s(\hat{\beta}_j) = \sqrt{\widehat{\boldsymbol{V}}_{\widehat{\beta}_j}} = \sqrt{\left[\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\right]_{jj}}.$$

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 $\hat{\theta}$ is the square root of (6.40)

$$s(\hat{\theta}) = \sqrt{\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}} = \sqrt{n^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} \widehat{\boldsymbol{R}}}.$$

When the justification is based on asymptotic theory we call $s(\hat{\beta}_j)$ or $s(\hat{\theta})$ an **asymptotic standard error** for $\hat{\beta}_j$ or $\hat{\theta}$. 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

$$\boldsymbol{R} = \begin{pmatrix} 100 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 100 \\ 20 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ -50/\beta_3 \\ 50\beta_2/\beta_3^2 \\ 0 \end{pmatrix},$$

respectively.

We use the subsample of married black women (all experience levels), which has 982 observations. The point estimates and standard errors are

$$\widehat{\log(Wage)} = \underset{(0.008)}{0.118}\ education + \underset{(0.006)}{0.016}\ experience - \underset{(0.012)}{0.022}\ experience^2/100 + \underset{(0.157)}{0.947}\ . \tag{6.41}$$

The standard errors are the square roots of the Horn-Horn-Duncan covariance matrix estimate

$$\overline{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} = \begin{pmatrix} 0.632 & 0.131 & -0.143 & -11.1 \\ 0.131 & 0.390 & -0.731 & -6.25 \\ -0.143 & -0.731 & 1.48 & 9.43 \\ -11.1 & -6.25 & 9.43 & 246 \end{pmatrix} \times 10^{-4}. \tag{6.42}$$

We calculate that

$$\begin{aligned} \widehat{\theta}_1 &= 100\widehat{\beta}_1 \\ &= 100 \times 0.118 \\ &= 11.8 \end{aligned}$$

$$\begin{aligned} s(\widehat{\theta}_1) &= \sqrt{100^2 \times 0.632 \times 10^{-4}} \\ &= 0.8 \end{aligned}$$

$$\begin{aligned} \widehat{\theta}_2 &= 100\widehat{\beta}_2 + 20\widehat{\beta}_3 \\ &= 100 \times 0.016 - 20 \times 0.022 \\ &= 1.16 \end{aligned}$$

$$\begin{aligned} s(\widehat{\theta}_2) &= \sqrt{\begin{pmatrix} 100 & 20 \end{pmatrix} \begin{pmatrix} 0.390 & -0.731 \\ -0.731 & 1.48 \end{pmatrix} \begin{pmatrix} 100 \\ 20 \end{pmatrix} \times 10^{-4}} \\ &= 0.55 \end{aligned}$$

$$\begin{aligned}\widehat{\theta}_3 &= -50\widehat{\beta}_2/\widehat{\beta}_3 \\ &= 50 \times 0.016/0.022 \\ &= 35.2\end{aligned}$$

$$\begin{aligned}s(\widehat{\theta}_3) &= \sqrt{\left(\begin{array}{cc} -50/\widehat{\beta}_3 & 50\widehat{\beta}_2/\widehat{\beta}_3^2 \end{array} \right) \left(\begin{array}{cc} 0.390 & -0.731 \\ -0.731 & 1.48 \end{array} \right) \left(\begin{array}{c} -50/\widehat{\beta}_3 \\ 50\widehat{\beta}_2/\widehat{\beta}_3^2 \end{array} \right) \times 10^{-4}} \\ &= 7.0\end{aligned}$$

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.

6.12 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

$$t_n(\theta) = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})}. \tag{6.43}$$

Different writers have called (6.43) 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 (6.43) we won't make this distinction, and will simply refer to $t_n(\theta)$ as a t-statistic or a t-ratio. We also often suppress the parameter dependence, writing it as t_n. The t-statistic is a simple function of the estimate, its standard error, and the parameter.

By Theorems 6.10.2 and 6.10.3, $\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathrm{N}(0, V_\theta)$ and $\widehat{V}_\theta \xrightarrow{p} V_\theta$. Thus

$$\begin{aligned}t_n(\theta) &= \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \\ &= \frac{\sqrt{n}\left(\widehat{\theta} - \theta\right)}{\sqrt{\widehat{V}_\theta}} \\ &\xrightarrow{d} \frac{\mathrm{N}(0, V_\theta)}{\sqrt{V_\theta}} \\ &= \mathrm{Z} \sim \mathrm{N}(0, 1).\end{aligned}$$

The last equality is by the property that linear scales of normal distributions are normal.

Thus the asymptotic distribution of the t-ratio $t_n(\theta)$ is the standard normal. Since this distribution does not depend on the parameters, we say that $t_n(\theta)$ is **asymptotically pivotal**. In special cases (such as the normal regression model, see Section 3.18), the statistic t_n has an exact t distribution, and is therefore exactly free of unknowns. In this case, we say that t_n is **exactly pivotal**. In general, however, pivotal statistics are unavailable and we must rely on asymptotically pivotal statistics.

As we will see in the next section, it is also useful to consider the distribution of the **absolute t-ratio** $|t_n(\theta)|$. Since $t_n(\theta) \xrightarrow{d} \mathrm{Z}$, the continuous mapping theorem yields $|t_n(\theta)| \xrightarrow{d} |\mathrm{Z}|$. 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

$$\begin{aligned}
\Pr(|\mathrm{Z}| \leq u) &= \Pr(-u \leq \mathrm{Z} \leq u) \\
&= \Pr(\mathrm{Z} \leq u) - \Pr(\mathrm{Z} < -u) \\
&= \Phi(u) - \Phi(-u) \\
&= 2\Phi(u) - 1 \\
&\overset{def}{=} \overline{\Phi}(u).
\end{aligned} \tag{6.44}$$

Theorem 6.12.1 *Under Assumptions 6.1.2 and 6.10.1,* $t_n(\theta) \xrightarrow{d} \mathrm{Z} \sim \mathrm{N}(0,1)$ *and* $|t_n(\theta)| \xrightarrow{d} |\mathrm{Z}|$.

The asymptotic normality of Theorem 6.12.1 is used to justify confidence intervals and tests for the parameters.

6.13 Confidence Intervals

The OLS estimate $\widehat{\boldsymbol{\beta}}$ is a **point estimate** for $\boldsymbol{\beta}$, meaning that $\widehat{\boldsymbol{\beta}}$ is a single value in $\mathbb{R}^k$. A broader concept is a **set estimate** C_n which is a collection of values in $\mathbb{R}^k$. When the parameter θ is real-valued then it is common to focus on intervals $C_n = [L_n, U_n]$ which is called an **interval estimate** for θ. The goal of an interval estimate C_n is to contain the true value, e.g. $\theta \in C_n$, with high probability.

The interval estimate C_n is a function of the data and hence is random. The **coverage probability** of the interval $C_n = [L_n, U_n]$ is $\Pr(\theta \in C_n)$. The randomness comes from C_n as the parameter θ is treated as fixed.

An interval estimates C_n 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%. C_n is called a $(1-\alpha)\%$ confidence interval if $\inf_\theta \Pr_\theta(\theta \in C_n) = 1 - \alpha$.

There is not a unique method to construct confidence intervals. One simple (yet silly) interval is

$$C_n = \begin{cases} \mathbb{R} & \text{with probability } 1-\alpha \\ \widehat{\theta} & \text{with probability } \alpha \end{cases}$$

By construction, if $\widehat{\theta}$ has a continuous distribution, $\Pr(\theta \in C_n) = 1 - \alpha$, so this confidence interval has perfect coverage, but C_n is uninformative about θ and is therefore not useful.

When we have an asymptotically normal parameter estimate $\widehat{\theta}$ with standard error $s(\widehat{\theta})$, the conventional confidence interval for θ takes the form

$$C_n = \left[\widehat{\theta} - c \cdot s(\widehat{\theta}), \quad \widehat{\theta} + c \cdot s(\widehat{\theta})\right] \tag{6.45}$$

where $c > 0$ is a pre-specified constant. This confidence interval is symmetric about the point estimate $\widehat{\theta}$, and its length is proportional to the standard error $s(\widehat{\theta})$.

Equivalently, C_n is the set of parameter values for θ such that the t-statistic $t_n(\theta)$ is smaller (in absolute value) than c, that is

$$C_n = \{\theta : |t_n(\theta)| \leq c\} = \left\{\theta : -c \leq \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \leq c\right\}.$$

The coverage probability of this confidence interval is

$$\Pr(\theta \in C_n) = \Pr(|t_n(\theta)| \leq c)$$

which is generally unknown. We can approximate the coverage probability by taking the asymptotic limit as $n \to \infty$. Since $|t_n(\theta)|$ is asymptotically $|\mathrm{Z}|$ (Theorem 6.12.1), it follows that as $n \to \infty$ that

$$\Pr(\theta \in C_n) \to \Pr(|\mathrm{Z}| \leq c) = \overline{\Phi}(c)$$

where $\overline{\Phi}(u)$ is given in (6.44). We call this the **asymptotic coverage probability**. Since the t-ratio is asymptotically pivotal, the asymptotic coverage probability is independent of the parameter θ, and is only a function of c.

As we mentioned before, an ideal confidence interval has a pre-specified probability coverage $1-\alpha$, typically 90% or 95%. This means selecting the constant c so that

$$\overline{\Phi}(c) = 1 - \alpha.$$

Effectively, this makes c a function of α, and can be backed out of a normal distribution table. For example, $\alpha = 0.05$ (a 95% interval) implies $c = 1.96$ and $\alpha = 0.1$ (a 90% interval) implies $c = 1.645$. Rounding 1.96 to 2, we obtain the most commonly used confidence interval in applied econometric practice

$$C_n = \left[\widehat{\theta} - 2s(\widehat{\theta}), \quad \widehat{\theta} + 2s(\widehat{\theta})\right]. \tag{6.46}$$

This is a useful rule-of thumb. This asymptotic 95% confidence interval C_n 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 6.13.1 *Under Assumptions 6.1.2 and 6.10.1, for C_n defined in (6.45),* $\Pr(\theta \in C_n) \longrightarrow \overline{\Phi}(c)$. *For* $c = 1.96$, $\Pr(\theta \in C_n) \longrightarrow 0.95$.

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 6.11 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 $35 \pm 1.28 \times 7 = [26, 44]$.

6.14 Regression Intervals

In the linear regression model the conditional mean of y_i given $\boldsymbol{x}_i = \boldsymbol{x}$ is

$$m(\boldsymbol{x}) = \mathbb{E}(y_i \mid \boldsymbol{x}_i = \boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}.$$

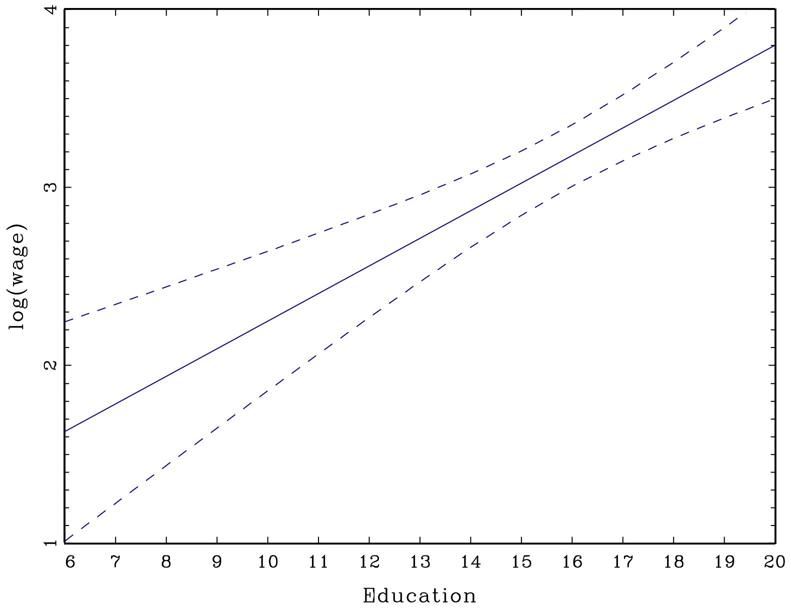


Figure 6.6: Wage on Education Regression Intervals

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 (3.11) 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 (4.35) 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 6.6. 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 (6.41) which includes experience and its square, with covariance matrix (6.42). 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 specific value. We select *education*=12. This only affects the level of the estimated regression, since

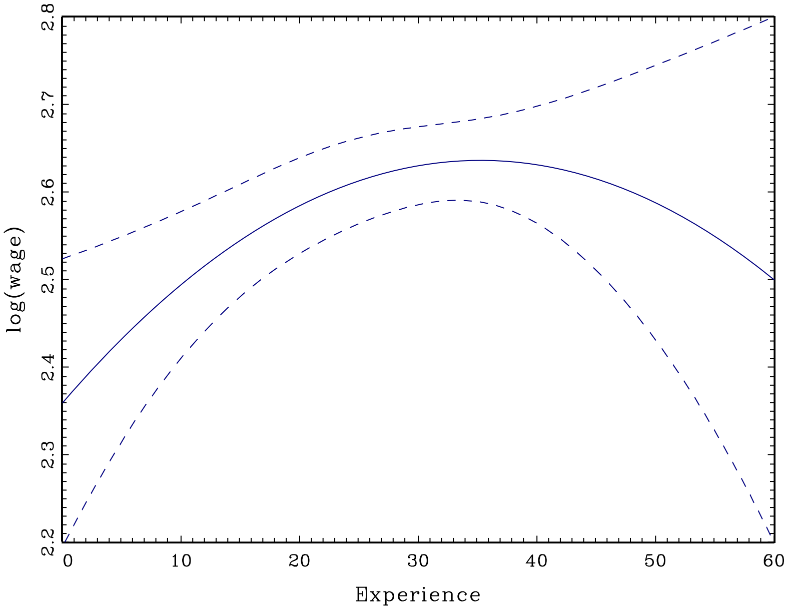


Figure 6.7: Wage on Experience Regression Intervals

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 6.7. 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 6.6 due to the nonlinear specification.

6.15 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.

The forecast error is $\hat{e}_{n+1} = y_{n+1} - \widehat{m}(\boldsymbol{x}) = e_{n+1} - \boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$. As the out-of-sample error e_{n+1} is independent of the in-sample estimate $\widehat{\boldsymbol{\beta}}$, this has conditional variance

$$\begin{aligned}\mathbb{E}\left(\hat{e}_{n+1}^2 | \boldsymbol{x}_{n+1} = \boldsymbol{x}\right) &= \mathbb{E}\left(e_{n+1}^2 - 2\boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)e_{n+1} + \boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\boldsymbol{x} | \boldsymbol{x}_{n+1} = \boldsymbol{x}\right) \\ &= \mathbb{E}\left(e_{n+1}^2 \mid \boldsymbol{x}_{n+1} = \boldsymbol{x}\right) + \boldsymbol{x}'\mathbb{E}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\boldsymbol{x} \\ &= \sigma^2(\boldsymbol{x}) + \boldsymbol{x}'\boldsymbol{V}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}.\end{aligned}$$

Under homoskedasticity $\mathbb{E}\left(e_{n+1}^2 \mid \boldsymbol{x}_{n+1}\right) = \sigma^2$, the natural estimate of this variance is $\hat{\sigma}^2 + \boldsymbol{x}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}$, so a standard error for the forecast is $\hat{s}(\boldsymbol{x}) = \sqrt{\hat{\sigma}^2 + \boldsymbol{x}'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}\boldsymbol{x}}$. 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

$$\left[\boldsymbol{x}'\widehat{\boldsymbol{\beta}} \pm 2\hat{s}(\boldsymbol{x})\right].$$

It is difficult, however, to fully justify this choice. It would be correct if we have a normal approximation to the ratio

$$\frac{e_{n+1} - \boldsymbol{x}'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{\hat{s}(\boldsymbol{x})}.$$

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 $\left[\boldsymbol{x}'\widehat{\boldsymbol{\beta}} \pm 2\hat{s}(\boldsymbol{x})\right]$ despite the lack of a convincing justification.

6.16 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

$$W_n(\theta) = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)'\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)'\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right). \qquad (6.47)$$

where $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} = n\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}$. When $q = 1$, then $W_n(\theta) = t_n(\theta)^2$ is the square of the t-ratio. When $q > 1$, $W_n(\theta)$ is typically called a **Wald statistic**. We are interested in its sampling distribution.

The asymptotic distribution of $W_n(\theta)$ is simple to derive given Theorem 6.10.2 and Theorem 6.10.3, which show that

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{Z} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

and

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{V}_{\boldsymbol{\theta}}.$$

It follows that

$$W_n(\boldsymbol{\theta}) = \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)'\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{Z}'\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\mathrm{Z} \qquad (6.48)$$

a quadratic in the normal random vector Z. Here we can appeal to a useful result from probability theory. (See Theorem B.9.3 in the Appendix.)

Theorem 6.16.1 *If* $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{A})$ *with* $\boldsymbol{A} > 0$, $q \times q$, *then* $\boldsymbol{Z}'\boldsymbol{A}^{-1}\boldsymbol{Z} \sim \chi^2_q$, *a chi-square random variable with* q *degrees of freedom.*

The asymptotic distribution in (6.48) takes exactly this form. Note that $\boldsymbol{V}_{\boldsymbol{\theta}} > 0$ since $\boldsymbol{R}$ is full rank under Assumption 6.10.1 It follows that $W_n(\boldsymbol{\theta})$ converges in distribution to a chi-square random variable.

Theorem 6.16.2 *Under Assumptions 6.1.2 and 6.10.1, as* $n \to \infty$,

$$W_n(\boldsymbol{\theta}) \xrightarrow{d} \chi^2_q.$$

Theorem 6.16.2 is used to justify multivariate confidence regions and mutivariate hypothesis tests.

6.17 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 $\widehat{\boldsymbol{V}}^0_{\boldsymbol{\theta}}$ defined in (6.39). This yields a homoskedastic Wald statistic

$$W^0_n(\theta) = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \left(\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\theta}}}\right)^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = n \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \left(\widehat{\boldsymbol{V}}^0_{\boldsymbol{\theta}}\right)^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right). \tag{6.49}$$

Under the additional assumption of conditional homoskedasticity, it has the same asymptotic distribution as $W_n(\theta)$.

Theorem 6.17.1 *Under Assumptions 6.1.2 and 6.10.1, and* $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, *as* $n \to \infty$,

$$W^0_n(\boldsymbol{\theta}) \xrightarrow{d} \chi^2_q.$$

6.18 Confidence Regions

A confidence region C_n is a set estimator for $\boldsymbol{\theta} \in \mathbb{R}^q$ when $q > 1$. A confidence region C_n 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 $\Pr(\boldsymbol{\theta} \in C_n) = 1 - \alpha$. 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 6.16.2, a good choice for a confidence region is the ellipse

$$C_n = \left\{\boldsymbol{\theta} : W_n(\boldsymbol{\theta}) \leq c_{1-\alpha}\right\}.$$

with $c_{1-\alpha}$ the $1 - \alpha$'th quantile of the χ^2_q distribution. (Thus $F_q(c_{1-\alpha}) = 1 - \alpha$.) These quantiles can be found from the χ^2_q critical value table.

Theorem 6.16.2 implies

$$\Pr\left(\boldsymbol{\theta} \in C_n\right) \to \Pr\left(\chi^2_q \leq c_{1-\alpha}\right) = 1 - \alpha$$

which shows that C_n has asymptotic coverage $(1-\alpha)\%$.

To illustrate the construction of a confidence region, consider the estimated regression (6.41) 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

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 20 \end{pmatrix} \widehat{\boldsymbol{\beta}} = \begin{pmatrix} 11.8 \\ 1.2 \end{pmatrix},$$

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

$$\begin{aligned} W_n(\boldsymbol{\theta}) &= \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \\ &= \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix}' \begin{pmatrix} 1.77 & -1.16 \\ -1.16 & 7.13 \end{pmatrix} \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix} \\ &= 1.77\,(11.8 - \theta_1)^2 - 2.32\,(11.8 - \theta_1)\,(1.2 - \theta_2) + 7.13\,(1.2 - \theta_2)^2. \end{aligned}$$

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 $W_n(\boldsymbol{\theta}) = 4.605$ which is displayed in Figure 6.8. Since the estimated correlation of the two coefficient estimates is modest (about 0.3) the region is modestly elliptical.

6.19 Semiparametric Efficiency in the Projection Model

In Section 4.6 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 5.13.2. 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.

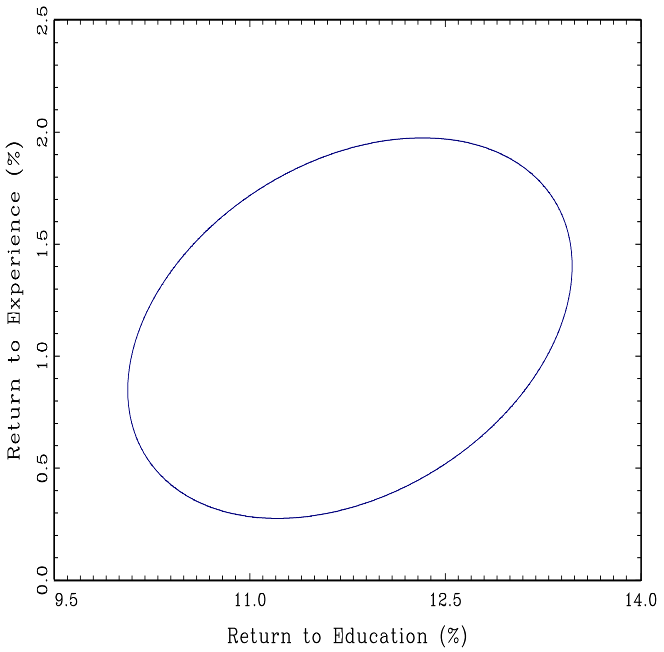


Figure 6.8: Confidence Region for Return to Experience and Return to Education

To see this, it is worth rephrasing Proposition 5.13.2 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 6.19.1 *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 6.19.1 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 $\boldsymbol{\beta} = \boldsymbol{Q}_{xx}^{-1}\boldsymbol{Q}_{xy} = \boldsymbol{g}(\boldsymbol{\mu})$ 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

$$\mathcal{L}_4(\boldsymbol{\beta}) = \left\{F : \mathbb{E}y^4 < \infty,\ \mathbb{E}\left\|\boldsymbol{x}\right\|^4 < \infty,\ \mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i' > 0\right\}.$$

Proposition 6.19.2 *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 6.19.1 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 $\hat{\sigma}^2$ for σ^2. To see this, note that we can write

$$\begin{aligned}\sigma^2 &= \mathbb{E}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\ &= \mathbb{E}y_i^2 - 2\mathbb{E}\left(y_i\boldsymbol{x}_i'\right)\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\boldsymbol{\beta} \\ &= Q_{yy} - \boldsymbol{Q}_{y\boldsymbol{x}}\boldsymbol{Q}_{\boldsymbol{xx}}^{-1}\boldsymbol{Q}_{\boldsymbol{x}y}\end{aligned}$$

which is a smooth function of the moments Q_{yy}, $\boldsymbol{Q}_{y\boldsymbol{x}}$ and $\boldsymbol{Q}_{\boldsymbol{xx}}$. Similarly the estimator $\hat{\sigma}^2$ equals

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2 \\ &= \widehat{Q}_{yy} - \widehat{\boldsymbol{Q}}_{y\boldsymbol{x}}\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1}\widehat{\boldsymbol{Q}}_{\boldsymbol{x}y}\end{aligned}$$

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 6.19.1 are satisfied. We conclude:

Proposition 6.19.3 *In the class of distributions $F \in \mathcal{L}_4(\boldsymbol{\beta})$, $\hat{\sigma}^2$ is semiparametrically efficient for σ^2.*

6.20 Semiparametric Efficiency in the Homoskedastic Regression Model*

In Section 6.19 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 $\boldsymbol{V}_{\boldsymbol{\beta}}^0 = \boldsymbol{Q}_{xx}^{-1}\sigma^2$. Therefore, as described in Section 5.13, 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

$$f(y, \boldsymbol{x} \mid \boldsymbol{\theta}) = f_1(y \mid \boldsymbol{x})\left(1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) f_2(\boldsymbol{x}). \tag{6.50}$$

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 $f_1(y \mid \boldsymbol{x})\left(1 + (y - \boldsymbol{x}'\boldsymbol{\beta})(\boldsymbol{x}'\boldsymbol{\theta})/\sigma^2\right)$. 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

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log f(y, \boldsymbol{x} \mid \boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(1 + \left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)\left(\boldsymbol{x}'\boldsymbol{\theta}\right)/\sigma^2\right) = \frac{\boldsymbol{x}\left(y - \boldsymbol{x}'\boldsymbol{\beta}\right)/\sigma^2}{1 + (y - \boldsymbol{x}'\boldsymbol{\beta})(\boldsymbol{x}'\boldsymbol{\theta})/\sigma^2}$$

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

$$\left(\mathbb{E}\left(\boldsymbol{s}\boldsymbol{s}'\right)\right)^{-1} = \left(\sigma^{-4}\mathbb{E}\left((\boldsymbol{x}e)(\boldsymbol{x}e)'\right)\right)^{-1} = \sigma^2 \boldsymbol{Q}_{xx}^{-1} = \boldsymbol{V}_{\beta}^0.$$

We have shown that there is a parametric submodel (6.50) 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 6.20.1 *In the homoskedastic regression model, the semiparametric variance bound for estimation of* $\boldsymbol{\beta}$ *is* $\boldsymbol{V}^0 = \sigma^2 \boldsymbol{Q}_{xx}^{-1}$ *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 6.20.1 states that OLS has the smallest asymptotic variance among all regular estimators. This is a much more powerful statement.

6.21 Uniformly Consistent Residuals*

It seems natural to view the residuals $\hat{e}_i$ 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

$$\begin{aligned} \hat{e}_i &= y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} \\ &= e_i + \boldsymbol{x}_i'\boldsymbol{\beta} - x_i'\widehat{\boldsymbol{\beta}} \\ &= e_i - \boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \end{aligned} \tag{6.51}$$

Since $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \xrightarrow{p} \mathbf{0}$ it seems reasonable to guess that $\hat{e}_i$ will be close to e_i if n is large.

We can bound the difference in (6.51) using the Schwarz inequality (A.15) to find

$$|\hat{e}_i - e_i| = \left|\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\right| \leq \|\boldsymbol{x}_i\| \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|. \tag{6.52}$$

To bound (6.52) we can use $\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| = O_p(n^{-1/2})$ from Theorem 6.3.2, 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 $|\hat{e}_i - e_i| \leq B\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| = O_p(n^{-1/2})$. However if the regressor does not have bounded support then we have to be more careful.

The key is Theorem 5.12.1 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 (6.52) we obtain

$$\begin{aligned} \max_{1\leq i\leq n}|\hat{e}_i - e_i| &\leq \max_{1\leq i\leq n}\|\boldsymbol{x}_i\|\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\| \\ &= o_p(n^{-1/2+1/r}). \end{aligned}$$

We have shown the following.

Theorem 6.21.1 *Under Assumption 6.1.2 and* $\mathbb{E}\|\boldsymbol{x}_i\|^r < \infty$, *then uniformly in* $1 \leq i \leq n$

$$\hat{e}_i = e_i + o_p(n^{-1/2+1/r}). \tag{6.53}$$

The rate of convergence in (6.53) depends on r. Assumption 6.1.2 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 5.12.1 we see that if $\mathbb{E}\exp(\boldsymbol{t}'\boldsymbol{x}_i) < \infty$ for all $\|\boldsymbol{t}\| < \infty$ then $\boldsymbol{x}_i = o_p(\log n)$ uniformly in i, and thus $\hat{e}_i = e_i + o_p\left(n^{-1/2}\log n\right)$.

We mentioned in Section 6.7 that there are multiple ways to prove the consistent of the covariance matrix estimator $\widehat{\boldsymbol{\Omega}}$. We now show that Theorem 6.21.1 provides one simple method to establish (6.31) and thus Theorem 6.7.1. Let $q_n = \max_{1\leq i\leq n}|\hat{e}_i - e_i| = o_p(n^{-1/4})$. Since

$$\hat{e}_i^2 - e_i^2 = 2e_i(\hat{e}_i - e_i) + (\hat{e}_i - e_i)^2,$$

then

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i' \left(\hat{e}_i^2 - e_i^2\right)\right\| &\leq \frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i\boldsymbol{x}_i'\right\| \left|\hat{e}_i^2 - e_i^2\right| \\ &\leq \frac{2}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i\right\|^2 \left|e_i\right| \left|\hat{e}_i - e_i\right| + \frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i\right\|^2 \left|\hat{e}_i - e_i\right|^2 \\ &\leq \frac{2}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i\right\|^2 \left|e_i\right| q_n + \frac{1}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i\right\|^2 q_n^2 \\ &\leq o_p(n^{-1/4}).\end{aligned}$$

6.22 Asymptotic Leverage*

Recall the definition of leverage from (3.21)

$$h_{ii} = \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i.$$

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 $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{Q}_{xx} > 0$ then by the CMT, $\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right) \xrightarrow{p} \lambda_{\min}\left(\boldsymbol{Q}_{xx}\right) > 0$. (The latter is positive since $\boldsymbol{Q}_{xx}$ is positive definite and thus all its eigenvalues are positive.) Then by the Quadratic Inequality (A.23)

$$\begin{aligned}h_{ii} &= \boldsymbol{x}_i' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \\ &\leq \lambda_{\max}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(\boldsymbol{x}_i'\boldsymbol{x}_i\right) \\ &= \left(\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)\right)^{-1} \frac{1}{n}\left\|\boldsymbol{x}_i\right\|^2 \\ &\leq \left(\lambda_{\min}\left(\boldsymbol{Q}_{xx}\right) + o_p(1)\right)^{-1} \frac{1}{n} \max_{1\leq i\leq n} \left\|\boldsymbol{x}_i\right\|^2. \qquad (6.54)\end{aligned}$$

Theorem 5.12.1 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 (6.54) is $o_p\left(n^{2/r-1}\right)$.

Theorem 6.22.1 *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 (6.22.1) 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 6.1 Take the model $y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i$ with $\mathbb{E}\boldsymbol{x}_i e_i = \boldsymbol{0}$. 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 6.2 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

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' + \lambda \boldsymbol{I}_k\right)^{-1} \left(\sum_{i=1}^n \boldsymbol{x}_i y_i\right) \tag{6.55}$$

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 6.3 For the ridge regression estimator (6.55), 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 6.4 Verify some of the calculations reported in Section 6.4. 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) $\mathbb{E}x_{1i}^2 = 1$

(c) $\mathbb{E}x_{1i}x_{2i} = \frac{1}{2}$

(d) $\mathbb{E}\left(e_i^2\right) = 1$

(e) $\mathbb{E}\left(x_{1i}^2 e_i^2\right) = 1$

(f) $\mathbb{E}\left(x_{1i}x_{2i}e_i^2\right) = \frac{7}{8}$.

Exercise 6.5 Show (6.19)-(6.22).

Exercise 6.6 The model is

$$\begin{aligned}
y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\
\mathbb{E}(\boldsymbol{x}_i e_i) &= \boldsymbol{0} \\
\boldsymbol{\Omega} &= \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' e_i^2\right).
\end{aligned}$$

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 6.7 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

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \\ y_i &= y_i^* + u_i \end{aligned}$$

where u_i is a measurement error satisfying

$$\begin{aligned} \mathbb{E}\left(\boldsymbol{x}_i u_i\right) &= \mathbf{0} \\ \mathbb{E}\left(y_i^* u_i\right) &= 0 \end{aligned}$$

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 6.8 Find the asymptotic distribution of $\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right)$ as $n \to \infty$.

Exercise 6.9 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 $x_i \in \mathbb{R}$. 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 6.10 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 $\widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}}$, based on a sample of size n. Let $\hat{\sigma}^2$ 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 $(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{V}}_{\widehat{\boldsymbol{\beta}}}, \hat{\sigma}^2)$, the residuals $\hat{\boldsymbol{e}}$, 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 6.11 As in Exercise 3.21, 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 $\hat{s}(\widehat{\theta})$ 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 7

Restricted Estimation

7.1 Introduction

In the linear projection model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= 0 \end{aligned}$$

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

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= 0 \end{aligned}$$

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

$$\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c} \tag{7.1}$$

where $\boldsymbol{R}$ is $k \times q$, $\text{rank}(\boldsymbol{R}) = q < k$ 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).

The constraint $\boldsymbol{\beta}_2 = \mathbf{0}$ discussed above is a special case of the constraint (7.1) with

$$\boldsymbol{R} = \left(\begin{array}{c} \mathbf{0} \\ \boldsymbol{I} \end{array} \right), \tag{7.2}$$

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 (7.1)? 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.

7.2 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

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}}\, SSE_n(\boldsymbol{\beta}) \tag{7.3}$$

where

$$SSE_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}. \tag{7.4}$$

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 (7.1) 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 "^" 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 (7.3) uses the technique of Lagrange multipliers. The problem (7.3) is equivalent to the minimization of the Lagrangian

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} SSE_n(\boldsymbol{\beta}) + \boldsymbol{\lambda}'\left(\boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{c}\right) \tag{7.5}$$

over $(\boldsymbol{\beta}, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is an $s \times 1$ vector of Lagrange multipliers. The first-order conditions for minimization of (7.5) are

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathcal{L}(\widetilde{\boldsymbol{\beta}}_{\text{cls}}, \widetilde{\boldsymbol{\lambda}}_{\text{cls}}) = -\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{X}'\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{\text{cls}} + \boldsymbol{R}\widetilde{\boldsymbol{\lambda}}_{\text{cls}} = \boldsymbol{0} \tag{7.6}$$

and

$$\frac{\partial}{\partial \boldsymbol{\lambda}} \mathcal{L}(\widetilde{\boldsymbol{\beta}}_{\text{cls}}, \widetilde{\boldsymbol{\lambda}}_{\text{cls}}) = \boldsymbol{R}'\widetilde{\boldsymbol{\beta}} - \boldsymbol{c} = \boldsymbol{0}. \tag{7.7}$$

Premultiplying (7.6) by $\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ we obtain

$$-\boldsymbol{R}'\widehat{\boldsymbol{\beta}} + \boldsymbol{R}'\widetilde{\boldsymbol{\beta}}_{\text{cls}} + \boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\widetilde{\boldsymbol{\lambda}}_{\text{cls}} = \boldsymbol{0} \tag{7.8}$$

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 (7.7) 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.8.)

Substituting this expression into (7.6) and solving for $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ we find the solution to the constrained minimization problem (7.3)

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \widehat{\boldsymbol{\beta}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left[\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right). \tag{7.9}$$

(See Exercise 7.4 to verify that (7.9) satisfies (7.1).)

This is a general formula for the CLS estimator. It also can be written as

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right). \tag{7.10}$$

Given $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ the residuals are

$$\tilde{e}_i = y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}_{\text{cls}}. \tag{7.11}$$

The moment estimator of σ^2 is

$$\tilde{\sigma}^2_{\text{cls}} = \frac{1}{n}\sum_{i=1}^n \tilde{e}_i^2.$$

A bias-corrected version of $\tilde{\sigma}^2_{\text{cls}}$ is

$$s^2_{\text{cls}} = \frac{1}{n-k+q}\sum_{i=1}^n \tilde{e}_i^2.$$

You can show (See Exercise 7.6) that in the homoskedastic linear regression model under (7.1),

$$\mathbb{E}\left(s^2_{\text{cls}} \mid \boldsymbol{X}\right) = \sigma^2 \tag{7.12}$$

so that s^2_{cls} is unbiased for σ^2.

7.3 Exclusion Restriction

While (7.9) 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

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \tag{7.13}$$

the exclusion restriction is $\boldsymbol{\beta}_2 = \boldsymbol{0}$, and the constrained equation is

$$y_i = \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_i. \tag{7.14}$$

In this setting the CLS estimator is OLS of y_i on x_{1i}. (See Exercise 7.1.) We can write this as

$$\widetilde{\boldsymbol{\beta}}_1 = \left(\sum_{i=1}^n \boldsymbol{x}_{1i}\boldsymbol{x}_{1i}'\right)^{-1}\left(\sum_{i=1}^n \boldsymbol{x}_{1i}y_i\right). \tag{7.15}$$

The CLS estimator of the entire vector $\boldsymbol{\beta}' = \left(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2'\right)$ is

$$\widetilde{\boldsymbol{\beta}} = \begin{pmatrix} \widetilde{\boldsymbol{\beta}}_1 \\ \boldsymbol{0} \end{pmatrix}. \tag{7.16}$$

It is not immediately obvious, but (7.9) and (7.16) are algebraically (and numerically) equivalent. To see this, the first component of (7.9) with (7.2) is

$$\widetilde{\boldsymbol{\beta}}_1 = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \end{pmatrix}\left[\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{Q}}_{xx}^{-1}\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{I} \end{pmatrix}\left[\begin{pmatrix} \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}\widehat{\boldsymbol{Q}}_{xx}^{-1}\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{I} \end{pmatrix}\right]^{-1}\begin{pmatrix} \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}\widehat{\boldsymbol{\beta}}\right].$$

Using (3.33) this equals

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}_1 &= \widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{Q}}^{12}\left(\widehat{\boldsymbol{Q}}^{22}\right)^{-1}\widehat{\boldsymbol{\beta}}_2 \\
&= \widehat{\boldsymbol{\beta}}_1 + \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{22\cdot 1}\widehat{\boldsymbol{\beta}}_2 \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{1y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{2y}\right) \\
&\quad + \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{22\cdot 1}\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\left(\widehat{\boldsymbol{Q}}_{2y} - \widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}\right) \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{1y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}\right) \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}\right)\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y} \\
&= \widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1y}
\end{aligned}$$

which is (7.16) as originally claimed.

7.4 Minimum Distance

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 $\boldsymbol{W}_n > 0$ define the quadratic criterion function

$$J_n(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \boldsymbol{W}_n \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{7.17}$$

This is a (squared) weighted Euclidean distance between $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$. $J_n(\boldsymbol{\beta})$ 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 $J_n(\boldsymbol{\beta})$ subject to the constraint (7.1), that is,

$$\widetilde{\boldsymbol{\beta}}_{\text{md}} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}}\ J_n(\boldsymbol{\beta}). \tag{7.18}$$

The CLS estimator is the special case when $\boldsymbol{W}_n = \widehat{\boldsymbol{Q}}_{xx}$, and we write this criterion function as

$$J_n^0(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \widehat{\boldsymbol{Q}}_{xx} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right). \tag{7.19}$$

To see the equality of CLS and minimum distance, rewrite the least-squares criterion as follows. Write the unconstrained least-squares fitted equation as $y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \hat{e}_i$ and substitute this equation into $SSE_n(\boldsymbol{\beta})$ to obtain

$$\begin{aligned}
SSE_n(\boldsymbol{\beta}) &= \sum_{i=1}^n \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\
&= \sum_{i=1}^n \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \hat{e}_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 \\
&= \sum_{i=1}^n \hat{e}_i^2 + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)' \left(\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i'\right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&= n\hat{\sigma}^2 + J_n^0(\boldsymbol{\beta}) \qquad (7.20)
\end{aligned}$$

where the third equality uses the fact that $\sum_{i=1}^n \boldsymbol{x}_i\hat{e}_i = 0$, and the last line uses $\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i' = n\widehat{\boldsymbol{Q}}_{xx}$. The expression (7.20) only depends on $\boldsymbol{\beta}$ through $J_n^0(\boldsymbol{\beta})$. Thus minimization of $SSE_n(\boldsymbol{\beta})$ and $J_n^0(\boldsymbol{\beta})$ are equivalent, and hence $\widetilde{\boldsymbol{\beta}}_{\text{md}} = \widetilde{\boldsymbol{\beta}}_{\text{cls}}$ when $\boldsymbol{W}_n = \widehat{\boldsymbol{Q}}_{xx}$.

We can solve for $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ explicitly by the method of Lagrange multipliers. The Lagrangian is

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2} J_n(\boldsymbol{\beta}, \boldsymbol{W}_n) + \boldsymbol{\lambda}'\left(\boldsymbol{R}'\boldsymbol{\beta} - \boldsymbol{c}\right)$$

which is minimized over $(\boldsymbol{\beta}, \boldsymbol{\lambda})$. The solution is

$$\widetilde{\boldsymbol{\lambda}}_{\text{md}} = n\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right) \tag{7.21}$$

$$\widetilde{\boldsymbol{\beta}}_{\text{md}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{W}_n^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right). \tag{7.22}$$

(See Exercise 7.7.) Comparing (7.22) with (7.10) we can see that $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ specializes to $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ when we set $\boldsymbol{W}_n = \widehat{\boldsymbol{Q}}_{xx}$.

An obvious question is which weight matrix $\boldsymbol{W}_n$ is best. We will address this question after we derive the asymptotic distribution for a general weight matrix.

7.5 Asymptotic Distribution

We first show that the class of minimum distance estimators are consistent for the population parameters when the constraints are valid.

Assumption 7.5.1 $\boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{c}$ *where* $\boldsymbol{R}$ *is* $k \times q$ *with* $\operatorname{rank}(\boldsymbol{R}) = q$.

Assumption 7.5.2 $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{W} > 0$.

Theorem 7.5.1 ***Consistency***
Under Assumptions 6.1.1, 7.5.1, and 7.5.2, $\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} \xrightarrow{p} \boldsymbol{\beta}$ *as* $n \to \infty$.

For a proof, see Exercise 7.8.

Theorem 7.5.1 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 7.5.2 ***Asymptotic Normality***
Under Assumptions 6.1.2, 7.5.1, and 7.5.2,

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right) \tag{7.23}$$

as $n \to \infty$*, where*

$$\begin{aligned} \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) = \boldsymbol{V}_{\boldsymbol{\beta}} - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}} \\ -\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}^{-1} \\ +\boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}^{-1} \end{aligned} \tag{7.24}$$

and $\boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{Q}_{xx}^{-1}\boldsymbol{\Omega}\boldsymbol{Q}_{xx}^{-1}$.

For a proof, see Exercise 7.9.

Theorem 7.5.2 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 7.5.3 ***Asymptotic Distribution of CLS Estimator***
Under Assumptions 6.1.2 and 7.5.1, 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

$$\begin{aligned} \boldsymbol{V}_{\text{cls}} &= \boldsymbol{V}_{\boldsymbol{\beta}} - \boldsymbol{Q}_{xx}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}} \\ &\quad - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}_{xx}^{-1} \\ &\quad + \boldsymbol{Q}_{xx}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}_{xx}^{-1} \end{aligned}$$

For a proof, see Exercise 7.10.

7.6 Efficient Minimum Distance Estimator

Theorem 7.5.2 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 7.6.1 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 $\boldsymbol{W}_n = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}$ the **efficient minimum distance estimator** and takes the form

$$\widetilde{\boldsymbol{\beta}}_{\text{emd}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right). \tag{7.25}$$

The asymptotic distribution of (7.25) can be deduced from Theorem 7.5.2. (See Exercises 7.11 and 7.12.)

Theorem 7.6.1 ***Efficient Minimum Distance Estimator***
Under Assumptions 6.1.2 and 7.5.1,

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\text{emd}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}^{*}\right)$$

as $n \to \infty$, *where*

$$\boldsymbol{V}_{\boldsymbol{\beta}}^{*} = \boldsymbol{V}_{\boldsymbol{\beta}} - \boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}. \tag{7.26}$$

Since

$$\boldsymbol{V}_{\boldsymbol{\beta}}^{*} \leq \boldsymbol{V}_{\boldsymbol{\beta}} \tag{7.27}$$

the estimator (7.25) has lower asymptotic variance than the unrestricted estimator. Furthermore, for any $\boldsymbol{W}$,

$$\boldsymbol{V}_{\boldsymbol{\beta}}^{*} \leq \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) \tag{7.28}$$

so (7.25) is asymptotically efficient in the class of minimum distance estimators.

Theorem 7.6.1 shows that the minimum distance estimator with the smallest asymptotic variance is (7.25). 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 $\boldsymbol{W} = \left(\boldsymbol{V}_{\boldsymbol{\beta}}^{0}\right)^{-1}$ 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 7.6.1 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 $\mathbb{E}\left(\boldsymbol{x}_{2i}e_i\right) = \boldsymbol{0}$. This information is relevant when the error is heteroskedastic and the excluded regressors are correlated with the included regressors.

Inequality (7.27) 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.

7.7 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 $\boldsymbol{\beta}_2 = \boldsymbol{0}$. We have introduced three estimators of $\boldsymbol{\beta}_1$. The first is unconstrained least-squares applied to (7.13), which can be written as

$$\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{1y\cdot 2}.$$

From Theorem 6.33 and equation (6.20) 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 7.5.3, but it is simpler to apply the CLT directly to show that

$$\operatorname{avar}(\widetilde{\boldsymbol{\beta}}_{1,\text{cls}}) = \boldsymbol{Q}_{11}^{-1}\boldsymbol{\Omega}_{11}\boldsymbol{Q}_{11}^{-1}. \tag{7.29}$$

The third estimator of $\boldsymbol{\beta}_1$ is the efficient minimum distance estimator. Applying (7.25), it equals

$$\widetilde{\boldsymbol{\beta}}_{1,\text{md}} = \widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{V}}_{12}\widehat{\boldsymbol{V}}_{22}^{-1}\widehat{\boldsymbol{\beta}}_2 \tag{7.30}$$

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 7.6.1 its asymptotic variance is

$$\operatorname{avar}(\widetilde{\boldsymbol{\beta}}_{1,\text{md}}) = \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}. \tag{7.31}$$

See Exercise 7.13 to verify equations (7.29), (7.30), and (7.31).

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

$$\operatorname{avar}(\widehat{\boldsymbol{\beta}}_1) = \sigma^2 \boldsymbol{Q}_{11\cdot 2}^{-1}$$

and

$$\operatorname{avar}(\widetilde{\boldsymbol{\beta}}_{1,\mathrm{cls}}) = \sigma^2 \boldsymbol{Q}_{11}^{-1}.$$

If $\boldsymbol{Q}_{12} = 0$ (so $\boldsymbol{x}_{1i}$ and $\boldsymbol{x}_{2i}$ are orthogonal) then these two variance matrices 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, $\widetilde{\boldsymbol{\beta}}_{1,\mathrm{cls}}$ 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 6.4. In that example, $Q_{11} = Q_{22} = 1$, $Q_{12} = \frac{1}{2}$, $\Omega_{11} = \Omega_{22} = 1$, and $\Omega_{12} = \frac{7}{8}$. We can calculate (see Exercise 7.14) that $\boldsymbol{Q}_{11\cdot 2} = \frac{3}{4}$ and

$$\operatorname{avar}(\widehat{\boldsymbol{\beta}}_1) = \frac{2}{3} \tag{7.32}$$

$$\operatorname{avar}(\widetilde{\boldsymbol{\beta}}_{1,\mathrm{cls}}) = 1 \tag{7.33}$$

$$\operatorname{avar}(\widetilde{\boldsymbol{\beta}}_{1,\mathrm{md}}) = \frac{5}{8}. \tag{7.34}$$

Thus the restricted least-squares estimator $\widetilde{\boldsymbol{\beta}}_{1,\mathrm{cls}}$ 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 (7.14) to be less efficient for estimation of $\boldsymbol{\beta}_1$ than least-squares applied to the long regression (7.13), 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.

7.8 Variance and Standard Error Estimation

The asymptotic covariance matrix (7.26) may be estimated by replacing $\boldsymbol{V}_{\boldsymbol{\beta}}$ with a consistent estimates such as $\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}$. This variance estimator is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{*} = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} - \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}. \tag{7.35}$$

We can calculate standard errors for any linear combination $\boldsymbol{h}'\widetilde{\boldsymbol{\beta}}$ 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

$$s(\boldsymbol{h}'\widetilde{\boldsymbol{\beta}}) = \left(n^{-1}\boldsymbol{h}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{*}\boldsymbol{h}\right)^{1/2}. \tag{7.36}$$

7.9 Misspecification

What are the consequences for a constrained estimator $\widetilde{\boldsymbol{\beta}}$ if the constraint (7.1) 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. (7.15)) is estimating a different projection coefficient than the long regression (e.g. (7.13)).

One mechanical answer is that we can use the formula (7.22) for the minimum distance estimator to find that

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} \xrightarrow{p} \boldsymbol{\beta}_{\mathrm{md}}^{*} = \boldsymbol{\beta} - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}(\boldsymbol{c}^{*} - \boldsymbol{c}). \tag{7.37}$$

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 $\boldsymbol{\beta}_{\mathrm{md}}^{*}$ 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

$$\boldsymbol{\beta}_{\mathrm{cls}}^{*} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}}\, \mathbb{E}\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2,$$

the best linear predictor subject to the constraint (7.1). The minimum distance estimator converges to

$$\boldsymbol{\beta}_{\mathrm{md}}^{*} = \underset{\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{c}}{\operatorname{argmin}}\, (\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\,\boldsymbol{W}\,(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

where $\boldsymbol{\beta}_0$ is the true coefficient. That is, $\boldsymbol{\beta}_{\mathrm{md}}^{*}$ is the coefficient vector satisfying (7.1) closest to the true value ın 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 $\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}$ has an asymptotic normal distribution. The trick is to define the pseudo-true value

$$\boldsymbol{\beta}_{n}^{*} = \boldsymbol{\beta} - \boldsymbol{W}_n^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}(\boldsymbol{c}^{*} - \boldsymbol{c}). \tag{7.38}$$

(Note that (7.37) and (7.38) are different!) Then

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} - \boldsymbol{\beta}_n^{*}\right) &= \sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) - \boldsymbol{W}_n^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}\sqrt{n}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}^{*}\right) \\
&= \left(\boldsymbol{I} - \boldsymbol{W}_n^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \\
&\xrightarrow{d} \left(\boldsymbol{I} - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right) \\
&= \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right).
\end{aligned} \tag{7.39}$$

In particular

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{emd}}-\boldsymbol{\beta}_n^*\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}^*\right).$$

This means that even when the constraint (7.1) is misspecified, the conventional covariance matrix estimator (7.35) and standard errors (7.36) 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

$$\boldsymbol{R}'\boldsymbol{\beta}_n=\boldsymbol{c}+\boldsymbol{\delta}n^{-1/2}. \tag{7.40}$$

Equation (7.40) specifies that $\boldsymbol{\beta}_n$ violates (7.1) 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 (7.40) **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 (7.1). 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

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_n\right)&=\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{x}_ie_i\right)\\&\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right).\end{aligned} \tag{7.41}$$

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 (7.40), $\boldsymbol{c}=\boldsymbol{R}'\boldsymbol{\beta}_n-\boldsymbol{\delta}n^{-1/2}$, so

$$\boldsymbol{R}'\widehat{\boldsymbol{\beta}}-\boldsymbol{c}=\boldsymbol{R}'\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_n\right)+\boldsymbol{\delta}n^{-1/2}$$

and

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}&=\widehat{\boldsymbol{\beta}}-\boldsymbol{W}_n^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}-\boldsymbol{c}\right)\\&=\widehat{\boldsymbol{\beta}}-\boldsymbol{W}_n^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_n\right)+\boldsymbol{W}_n^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta}n^{-1/2}.\end{aligned}$$

It follows that

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}-\boldsymbol{\beta}_n\right)=&\left(\boldsymbol{I}-\boldsymbol{W}_n^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}_n\right)\\&+\boldsymbol{W}_n^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}_n^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta}.\end{aligned}$$

The first term is asymptotically normal (from 7.41)). The second term converges in probability to a constant. This is because the $n^{-1/2}$ local scaling in (7.40) 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

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}-\boldsymbol{\beta}_n\right)&\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\beta}}\right)+\boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta}\\&=\mathrm{N}\left(\boldsymbol{\delta}^*,\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right)\end{aligned} \tag{7.42}$$

where

$$\boldsymbol{\delta}^* = \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{\delta}.$$

The asymptotic distribution (7.42) is an approximation of the sampling distribution of the restricted estimator under misspecification. The distribution (7.42) contains an asymptotic bias component $\boldsymbol{\delta}^*$. The approximation is not fundamentally different from (7.39) – they both have the same asymptotic variances, and both reflect the bias due to misspecification. The difference is that (7.39) puts the bias on the left-side of the convergence arrow, while (7.42) has the bias on the right-side. There is no substantive difference between the two, but (7.42) is more convenient for some purposes, such as the analysis of the power of tests, as we will explore in the next chapter.

7.10 Nonlinear Constraints

In some cases it is desirable to impose nonlinear constraints on the parameter vector $\boldsymbol{\beta}$. They can be written as

$$\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{0} \tag{7.43}$$

where $\boldsymbol{r} : \mathbb{R}^k \to \mathbb{R}^q$. This includes the linear constraints (7.1) as a special case. An example of (7.43) which cannot be written as (7.1) is $\beta_1\beta_2 = 1$, which is (7.43) with $r(\boldsymbol{\beta}) = \beta_1\beta_2 - 1$.

The constrained least-squares and minimum distance estimators of $\boldsymbol{\beta}$ subject to (7.43) solve the minimization problems

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})=\boldsymbol{0}}{\operatorname{argmin}}\, SSE_n(\boldsymbol{\beta}) \tag{7.44}$$

$$\widetilde{\boldsymbol{\beta}}_{\text{md}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})=\boldsymbol{0}}{\operatorname{argmin}}\; J_n(\boldsymbol{\beta}) \tag{7.45}$$

where $SSE_n(\boldsymbol{\beta})$ and $J_n(\boldsymbol{\beta})$ are defined in (7.4) and (7.17), respectively. The solutions minimize the Lagrangians

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{1}{2}SSE_n(\boldsymbol{\beta}) + \boldsymbol{\lambda}'\boldsymbol{r}(\boldsymbol{\beta}) \tag{7.46}$$

or

$$\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}) = \frac{1}{2}J_n(\boldsymbol{\beta}) + \boldsymbol{\lambda}'\boldsymbol{r}(\boldsymbol{\beta}) \tag{7.47}$$

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 (7.44) and (7.45) are known as **constrained optimization** methods, and are available in programming languages including Matlab, Gauss and R.

Assumption 7.10.1 $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{0}$ *with* $\operatorname{rank}(\boldsymbol{R}) = q$, *where* $\boldsymbol{R} = \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{r}(\boldsymbol{\beta})'$.

The asymptotic distribution is a simple generalization of the case of a linear constraint, but the proof is more delicate.

Theorem 7.10.1 *Under Assumptions 6.1.2, 7.10.1, and 7.5.2, for* $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_{\mathrm{md}}$ *and* $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_{\mathrm{cls}}$ *defined in (7.44) and (7.45),*

$$\sqrt{n}\left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})\right)$$

as $n \to \infty$, *where* $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})$ *is defined in (7.24). For* $\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}}$, $\boldsymbol{W} = \boldsymbol{Q}_{xx}$ *and* $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) = \boldsymbol{V}_{\mathrm{cls}}$ *as defined in Theorem 7.5.3.* $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})$ *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

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{r}(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}})'. \tag{7.48}$$

Standard errors for the elements of $\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}$ are the square roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widetilde{\boldsymbol{\beta}}}^{*} = n^{-1}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{*}$.

7.11 Inequality Restrictions

Inequality constraints on the parameter vector $\boldsymbol{\beta}$ take the form

$$\boldsymbol{r}(\boldsymbol{\beta}) \geq \mathbf{0} \tag{7.49}$$

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

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})\geq\mathbf{0}}{\mathrm{argmin}}\, SSE_n(\boldsymbol{\beta}) \tag{7.50}$$

and

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{md}} = \underset{\boldsymbol{r}(\boldsymbol{\beta})\geq\mathbf{0}}{\mathrm{argmin}}\; J_n\left(\boldsymbol{\beta}\right). \tag{7.51}$$

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 (7.50) and (7.51) 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 $\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \mathrm{Z} = \mathrm{N}\left(0, V_\beta\right)$ and $\beta = 0$, then the constrained estimator $\tilde{\beta} = \max[\hat{\beta}, 0]$ will have the asymptotic distribution $\sqrt{n}\tilde{\beta} \xrightarrow{d} \max[\mathrm{Z}, 0]$, a "half-normal" distribution.

7.12 Constrained MLE

Recall that the log-likelihood function (3.44) for the normal regression model is

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}SSE_n(\boldsymbol{\beta}).$$

The constrained maximum likelihood estimator (CMLE) $(\widehat{\boldsymbol{\beta}}_{\text{cmle}}, \hat{\sigma}^2_{\text{cmle}})$ maximizes $\log L(\boldsymbol{\beta}, \sigma^2)$ subject to the constraint (7.43) Since $\log L(\boldsymbol{\beta}, \sigma^2)$ is a function of $\boldsymbol{\beta}$ only through the sum of squared errors $SSE_n(\boldsymbol{\beta})$, maximizing the likelihood is identical to minimizing $SSE_n(\boldsymbol{\beta})$. Hence $\widehat{\boldsymbol{\beta}}_{\text{cmle}} = \widehat{\boldsymbol{\beta}}_{\text{cls}}$ and $\hat{\sigma}^2_{\text{cmle}} = \hat{\sigma}^2_{\text{cls}}$.

7.13 Technical Proofs*

Proof of Theorem 7.6.1, Equation (7.28). 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

$$\begin{aligned}\boldsymbol{C}'\boldsymbol{V}_\beta^*\boldsymbol{C} &= \left[\begin{array}{cc}\boldsymbol{R}'\boldsymbol{V}_\beta^*\boldsymbol{R} & \boldsymbol{R}'\boldsymbol{V}_\beta^*\boldsymbol{R}_\perp \\ \boldsymbol{R}_\perp'\boldsymbol{V}_\beta^*\boldsymbol{R} & \boldsymbol{R}_\perp'\boldsymbol{V}_\beta^*\boldsymbol{R}_\perp\end{array}\right] \\ &= \left[\begin{array}{cc}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}_\perp'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}_\perp\end{array}\right]\end{aligned}$$

and

$$\begin{aligned}&\boldsymbol{C}'\boldsymbol{V}_\beta(\boldsymbol{W})\boldsymbol{C} \\ &= \left[\begin{array}{cc}\boldsymbol{R}'\boldsymbol{V}_\beta^*(\boldsymbol{W})\boldsymbol{R} & \boldsymbol{R}'\boldsymbol{V}_\beta^*(\boldsymbol{W})\boldsymbol{R}_\perp \\ \boldsymbol{R}_\perp'\boldsymbol{V}_\beta^*(\boldsymbol{W})\boldsymbol{R} & \boldsymbol{R}_\perp'\boldsymbol{V}_\beta^*(\boldsymbol{W})\boldsymbol{R}_\perp\end{array}\right] \\ &= \left[\begin{array}{cc}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}_\perp'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}_\perp + \boldsymbol{R}_\perp'\boldsymbol{W}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}\right)^{-\mathbf{1}}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}\right)^{-\mathbf{1}}\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}_\perp\end{array}\right].\end{aligned}$$

Thus

$$\begin{aligned}&\boldsymbol{C}'\left(\boldsymbol{V}_\beta(\boldsymbol{W}) - \boldsymbol{V}_\beta^*\right)\boldsymbol{C} \\ &= \boldsymbol{C}'\boldsymbol{V}_\beta(\boldsymbol{W})\boldsymbol{C} - \boldsymbol{C}'\boldsymbol{V}_\beta^*\boldsymbol{C} \\ &= \left[\begin{array}{cc}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}_\perp'\boldsymbol{W}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}\right)^{-\mathbf{1}}\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}\right)^{-\mathbf{1}}\boldsymbol{R}'\boldsymbol{W}\boldsymbol{R}_\perp\end{array}\right] \\ &\geq 0\end{aligned}$$

Since $\boldsymbol{C}$ is invertible it follows that $\boldsymbol{V}_\beta(\boldsymbol{W}) - \boldsymbol{V}_\beta^* \geq 0$ which is (7.28). ■

Proof of Theorem 7.10.1. We show the result for the minimum distance estimator $\widetilde{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}}_{\text{md}}$, 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

$$\boldsymbol{r}_j(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{r}_j(\boldsymbol{\beta}) + \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}_j(\boldsymbol{\beta}_j^*)' \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \right). \tag{7.52}$$

Let $\boldsymbol{R}_n^*$ be the $k \times q$ matrix

$$\boldsymbol{R}_n^* = \left[\begin{array}{cccc} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}_1(\boldsymbol{\beta}_1^*) & \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}_2(\boldsymbol{\beta}_2^*) & \cdots & \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{r}_q(\boldsymbol{\beta}_q^*) \end{array} \right].$$

Since $\widetilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ it follows that $\boldsymbol{\beta}_j^* \xrightarrow{p} \boldsymbol{\beta}$, and by the CMT, $\boldsymbol{R}_n^* \xrightarrow{p} \boldsymbol{R}$. Stacking the (7.52), we obtain

$$\boldsymbol{r}(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{r}(\boldsymbol{\beta}) + \boldsymbol{R}_n^{*\prime} \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \right).$$

Since $\boldsymbol{r}(\widetilde{\boldsymbol{\beta}}) = \boldsymbol{0}$ by construction and $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{0}$ by Assumption 7.5.1, this implies

$$\boldsymbol{0} = \boldsymbol{R}_n^{*\prime} \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \right). \tag{7.53}$$

The first-order condition for (7.47) is

$$\boldsymbol{W}_n \left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}} \right) = \widehat{\boldsymbol{R}} \widetilde{\boldsymbol{\lambda}}.$$

where $\widehat{\boldsymbol{R}}$ is defined in (7.48).

Premultiplying by $\boldsymbol{R}^{*\prime} \boldsymbol{W}_n^{-1}$, inverting, and using (7.53), we find

$$\widetilde{\boldsymbol{\lambda}} = \left(\boldsymbol{R}_n^{*\prime} \boldsymbol{W}_n^{-1} \widehat{\boldsymbol{R}} \right)^{-1} \boldsymbol{R}_n^{*\prime} \left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}} \right) = \left(\boldsymbol{R}_n^{*\prime} \boldsymbol{W}_n^{-1} \widehat{\boldsymbol{R}} \right)^{-1} \boldsymbol{R}_n^{*\prime} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right).$$

Thus

$$\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\boldsymbol{I} - \boldsymbol{W}_n^{-1} \widehat{\boldsymbol{R}} \left(\boldsymbol{R}_n^{*\prime} \boldsymbol{W}_n^{-1} \widehat{\boldsymbol{R}} \right)^{-1} \boldsymbol{R}_n^{*\prime} \right) \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right). \tag{7.54}$$

From Theorem 6.3.2 and Theorem 6.7.1 we find

$$\begin{aligned}
\sqrt{n} \left(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) &= \left(\boldsymbol{I} - \boldsymbol{W}_n^{-1} \widehat{\boldsymbol{R}} \left(\boldsymbol{R}_n^{*\prime} \boldsymbol{W}_n^{-1} \widetilde{\boldsymbol{R}} \right)^{-1} \boldsymbol{R}_n^{*\prime} \right) \sqrt{n} \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right) \\
&\xrightarrow{d} \left(\boldsymbol{I} - \boldsymbol{W}^{-1} \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{W}^{-1} \boldsymbol{R} \right)^{-1} \boldsymbol{R}' \right) \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}} \right) \\
&= \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W}) \right).
\end{aligned}$$

■

Exercises

Exercise 7.1 In the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, show directly from definition (7.3) that the CLS estimate of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ subject to the constraint that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ is the OLS regression of $\boldsymbol{y}$ on $\boldsymbol{X}_1$.

Exercise 7.2 In the model $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{e}$, show directly from definition (7.3) 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 7.3 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 7.4 Verify that for $\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ defined in (7.9) that $\boldsymbol{R}'\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \boldsymbol{c}$.

Exercise 7.5 Let $\widetilde{\boldsymbol{e}}$ be the vector of constrained least-squares residuals (7.11). Show that under (7.1),

(a) $\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c} = \boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}$

(b) $\widetilde{\boldsymbol{\beta}}_{\text{cls}} - \boldsymbol{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}$

(c) $\widetilde{\boldsymbol{e}} = (\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{A})\,\boldsymbol{e}$ for $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ and some matrix $\boldsymbol{A}$ (find this matrix $\boldsymbol{A}$).

(d) Show that $\boldsymbol{A}$ is symmetric and idempotent, $\operatorname{tr}\boldsymbol{A} = q$, and $\boldsymbol{P}\boldsymbol{A} = \boldsymbol{A}$.

Exercise 7.6 Show (7.12), that is, $\mathbb{E}\left(s^2_{\text{cls}} \mid \boldsymbol{X}\right) = \sigma^2$, under the assumptions of the homoskedastic regression model and (7.1).

Hint: Use the results of Exercise 7.5

Exercise 7.7 Verify (7.21) and (7.22), and that the minimum distance estimator $\widetilde{\boldsymbol{\beta}}_{\text{md}}$ with $\boldsymbol{W}_n = \widehat{\boldsymbol{Q}}_{xx}$ equals the CLS estimator.

Exercise 7.8 Prove Theorem 7.5.1.

Exercise 7.9 Prove Theorem 7.5.2.

Exercise 7.10 Prove Theorem 7.5.3. (Hint: Use that CLS is a special case of Theorem 7.5.2.)

Exercise 7.11 Verify that (7.26) is $\boldsymbol{V}_{\boldsymbol{\beta}}(\boldsymbol{W})$ with $\boldsymbol{W} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}$.

Exercise 7.12 Prove (7.27). Hint: Use (7.26).

Exercise 7.13 Verify (7.29), (7.30) and (7.31)

Exercise 7.14 Verify (7.32), (7.33), and (7.34).

Exercise 7.15 As in Exercise 6.11 and 3.21, 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).

Chapter 8

Hypothesis Testing

8.1 Hypotheses

In Chapter 7 we discussed estimation subject to restrictions, including linear restrictions (7.1), nonlinear restrictions (7.43), and inequality restrictions (7.49). 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

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 \tag{8.1}$$

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 (8.1) 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 (8.1) as they are the most common and relatively simple to handle.

The hypothesis to be tested is called the null hypothesis.

Definition 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 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 $\mathbb{H}_0 : \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$. 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 zsero? 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?

8.2 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 can be 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**

$$T_n = T_n\left(\left(y_1, \boldsymbol{x}_1\right), ..., \left(y_n, \boldsymbol{x}_n\right)\right)$$

relative to a **critical value** c. The hypothesis test then consists of the decision rule

1. Accept $\mathbb{H}_0$ if $T_n \leq c$.

2. Reject $\mathbb{H}_0$ if $T_n > c$.

A test statistic T_n 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

$$t_n = |t_n(\theta_0)| \tag{8.2}$$

where

$$t_n(\theta) = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \tag{8.3}$$

is the t-statistic from (6.43), $\widehat{\theta}$ is a point estimate, and $s(\widehat{\theta})$ its standard error. t_n 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 $\mathbb{H}_0 : \theta = 1$.

8.3 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

$$\Pr\left(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ true}\right) = \Pr\left(T_n > c \mid \mathbb{H}_0 \text{ true}\right). \tag{8.4}$$

The finite sample **size** of the test is defined as the supremum of (8.4) 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 6, in typical econometric models the exact sampling distributions of estimators and test statistics are unknown and hence we cannot explicitly calculate (8.4). 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

$$T_n \xrightarrow{d} T \tag{8.5}$$

as $n \to \infty$ for some continuously-distributed random variable T. This is not a substantive restriction, as most conventional econometric tests satisfy (8.5). Let $G(u) = \Pr\left(T \leq u\right)$ denote the distribution of T. We call T (or G) the **asymptotic null distribution**.

It is generally desirable to design test statistics T_n whose asymptotic null distribution G is known and does not depend on unknown parameters. In this case we say that the statistic T_n is **asymptotically pivotal**.

For example, if $T_n = t_n$, the absolute t-statistic from (8.2), then we know from Theorem 6.12.1 that if $\theta = \theta_0$ (that is, the null hypothesis holds), then $t_n \xrightarrow{d} |\mathrm{Z}|$ as $n \to \infty$ where $\mathrm{Z} \sim \mathrm{N}(0,1)$. This means that $G(u) = \Pr\left(T \leq |\mathrm{Z}|\right) = \overline{\Phi}(u)$, the symmetrized normal distribution function defined in (6.44). 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:

$$\begin{aligned}\lim_{n\to\infty} \Pr\left(T_n > c \mid \mathbb{H}_0 \text{ true}\right) &= \Pr\left(T > c\right) \\ &= 1 - G(c).\end{aligned}$$

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 $1 - \overline{\Phi}(c)$.

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 $(1-\alpha)^{th}$ 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 $\overline{\Phi}(1.96) = 0.95$, it follows that the 5% asymptotic critical value for the absolute t-statistic is $c = 1.96$.

8.4 t tests

As we mentioned earlier, the most common test of the one-dimensional hypothesis

$$\mathbb{H}_0 : \theta = \theta_0 \tag{8.6}$$

against the alternative

$$\mathbb{H}_1 : \theta \neq \theta_0 \tag{8.7}$$

is the absolute value of the t-statistic (8.3). We now formally state its asymptotic null distribution, which is a simple application of Theorem 6.12.1.

Theorem 8.4.1 *Under Assumptions 6.1.2, 6.10.1, and* $\mathbb{H}_0 : \theta = \theta_0$,

$$t_n(\theta_0) \xrightarrow{d} \mathrm{Z}.$$

For c *satisfying* $\alpha = 2\left(1 - \Phi(c)\right)$,

$$\Pr\left(|t_n(\theta_0)| > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ *if* $|t_n(\theta_0)| > c$*" has asymptotic size* α.

The theorem shows that asymptotic critical values can be taken from the normal distribution table.

The alternative hypothesis (8.7) is sometimes called a "two-sided" alternative. In contrast, sometimes we are interested in testing for one-sided alternatives such as

$$\mathbb{H}_1 : \theta > \theta_0 \tag{8.8}$$

or

$$\mathbb{H}_1 : \theta < \theta_0. \tag{8.9}$$

Tests of (8.6) against (8.8) or (8.9) are based on the signed t-statistic $t_n = t_n(\theta_0)$. The hypothesis (8.6) is rejected in favor of (8.8) if $t_n > c$ where c satisfies $\alpha = 1 - \Phi(c)$. Negative values of t_n are not taken as evidence against $\mathbb{H}_0$, as point estimates $\widehat{\theta}$ less than θ_0 do not point to (8.8). 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 $\alpha = 1.645$. Thus, we reject (8.6) in favor of (8.8) if $t_n > 1.645$.

Conversely, tests of (8.6) against (8.9) reject $\mathbb{H}_0$ for negative t-statistics, e.g. if $t_n \leq -c$. For this alternative large positive values of t_n are not evidence against $\mathbb{H}_0$. An asymptotic 5% test rejects if $t_n < -1.645$.

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 (8.6) against (8.8) makes sense. If the restriction $\theta \geq \theta_0$ is not known *a priori*, then imposing this restriction to test (8.6) against (8.8) 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.

8.5 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:

$$\pi_n(\boldsymbol{\theta}) = \Pr\left(\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_1 \text{ true}\right) = \Pr\left(T_n > c \mid \mathbb{H}_1 \text{ true}\right).$$

We call $\pi_n(\boldsymbol{\theta})$ 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_n , 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 trivial 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.

8.6 Statistical Significance

Testing requires a pre-selected choice of significance level α, yet there is no objective scientific basis for choice of α. Never-the-less, 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 that it is **statistically insignificant**. It is helpful to remember that this is simply a compact way of saying "Using the statistic T_n, 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 $0.095/0.020 = 4.75$, 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 $0.022/0.020 = 1.10$, 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".

8.7 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 $0.180/0.008 = 22.5$ and $0.016/0.008 = 2.0$, respectively. Both exceed the asymptotic 5% critical value of 1.96, so we reject the hypothesis for both men and women. But the statistic for men is exceptionally high, and that for women is only slightly above the critical value. Suppose in contrast that the t-statistic had been 1.9, which is less than the critical value. This would lead to the decision "Accept $\mathbb{H}_0$" rather than "Reject $\mathbb{H}_0$". Should we really be making a different decision if the t-statistic is 1.9 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_n 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**

$$p_n = 1 - G(T_n).$$

Since the distribution function G is monotonically increasing, the p-value is a monotonically decreasing function of T_n and is an equivalent test statistic. Instead of rejecting $\mathbb{H}_0$ at the significance level α if $T_n > c$, we can reject $\mathbb{H}_0$ if $p_n < \alpha$. Thus it is sufficient to report p_n, and let the reader decide.

In is instructive to interpret p_n as the **marginal significance level**: the largest value of α for which the test T_n "rejects" the null hypothesis. That is, $p_n = 0.11$ means that T_n 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 $T_n \xrightarrow{d} T$ under $\mathbb{H}_0$, then $p_n = 1 - G(T_n) \xrightarrow{d} 1 - G(T)$, which has the distribution

$$\begin{aligned}
\Pr\left(1 - G(T) \le u\right) &= \Pr\left(1 - u \le G(T)\right) \\
&= 1 - \Pr\left(T \le G^{-1}(1-u)\right) \\
&= 1 - G\left(G^{-1}(1-u)\right) \\
&= 1 - (1-u) \\
&= u,
\end{aligned}$$

which is the uniform distribution on $[0,1]$. (This calculation assume that $G(u)$ is strictly increasing which is true for conventional asymptotic distributions such as the normal.) Thus $p_n \xrightarrow{d} \mathrm{U}[0,1]$. This means that the "unusualness" of p_n is easier to interpret than the "unusualness" of T_n.

An important caveat is that the p-value p_n should not be interpreted as the probability that either hypothesis is true. For example, a common mis-interpretation is that p_n is the probability "that the null hypothesis is false." This is incorrect. Rather, p_n is a measure of the strength of information against the null hypothesis.

Returing 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 highly unlikely to observe a t-statistic as large as 22.5 when the true value of the coefficient is zero, and thus we can reject that the true value 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 0.046. In this context it is typical to say that the test is "marginally significant", meaning that the test statistic is close to the asymptotic 5% critical value.

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 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_n rather than simply the test statistic T_n, 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_n with asymptotic distribution $T_n \xrightarrow{d} T$ under $\mathbb{H}_0$.

3. Set the asymptotic critical value c so that $1 - G(c) = \alpha$, where G is the distribution function of T.

4. Calculate the asymptotic p-value $p_n = 1 - G(T_n)$.

5. Reject $\mathbb{H}_0$ if $T_n > c$, or equivalently $p_n < \alpha$.

6. Accept $\mathbb{H}_0$ if $T_n \leq c$, or equivalently $p_n \geq \alpha$.

7. Report p_n to summarize the evidence concerning $\mathbb{H}_0$ versus $\mathbb{H}_1$.

8.8 t-ratios and the Abuse of Testing

In Section 4.15, we argued that a good applied practice is to report coefficient estimates $\hat{\theta}$ and standard errors $s(\hat{\theta})$ for all coefficients of interest in estimated models. With $\hat{\theta}$ and $s(\hat{\theta})$ the reader can easily construct confidence intervals $[\hat{\theta} \pm 2s(\hat{\theta})]$ and t-statistics $\left(\hat{\theta} - \theta_0\right)/s(\hat{\theta})$ for hypotheses of interest.

Some applied papers (especially older ones) instead report t-ratios $t_n = \hat{\theta}/s(\hat{\theta})$ 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. $s(\hat{\theta}) = \hat{\theta}/t_n$), 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.

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 "marginally significant" for women.

Now, let's construct asymptotic confidence intervals for the coefficients. The one for men is [0.16, 0.20] and that for women is [0.00, 0.03]. This shows that average wages for married men are about 16-20% higher than for unmarried men, which is very 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.

8.9 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 written in the form (8.1). 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 $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$. 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 (6.47) evaluated at the null hypothesis

$$W_n = W_n(\boldsymbol{\theta}_0) = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \tag{8.10}$$

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_n alternatively as

$$W_n = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$$

using the asymptotic variance estimate $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}$, or we can write it directly as a function of $\widehat{\boldsymbol{\beta}}$ as

$$W_n = \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}\right)^{-1} \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right). \tag{8.11}$$

Also, when $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{R}'\boldsymbol{\beta}$ is a linear function of $\boldsymbol{\beta}$, then the Wald statistic simplifies to

$$W_n = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\boldsymbol{R}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right).$$

The Wald statistic W_n is a weighted Euclidean measure of the length of the vector $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$. When $q = 1$ then $W_n = t_n^2$, the square of the t-statistic, so hypothesis tests based on W_n and $|t_n|$ are equivalent. The Wald statistic (8.10) 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 6.16.2, when $\boldsymbol{\beta}$ satisfies $\boldsymbol{r}(\boldsymbol{\beta}) = \boldsymbol{\theta}_0$ then $W_n \xrightarrow{d} \chi_q^2$, a chi-square random variable with q degrees of freedom. Let $G_q(u)$ denote the χ_q^2 distribution function. For a given significance level α, the asymptotic critical value c satisfies $\alpha = 1 - G_q(c)$ and can be found from the chi-square distribution table. For example, the 5% critical values for $q = 1$, $q = 2$, and $q = 3$ are 3.84, 5.99, and 7.82, respectively. An asymptotic test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $W_n > c$. As with t-tests, it is conventional to describe a Wald test as "significant" if W_n exceeds the 5% asymptotic critical value.

Theorem 8.9.1 *Under Assumptions 6.1.2 and 6.10.1, and* $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, *then*

$$W_n \xrightarrow{d} \chi_q^2,$$

and for c *satisfying* $\alpha = 1 - G_q(c)$,

$$\Pr\left(W_n > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ *if* $W_n > c$*" has asymptotic size* α.

Notice that the asymptotic distribution in Theorem 8.9.1 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_n is $p_n = 1 - G_q(W_n)$, 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 $W_n = 11.2$, it is difficult for a reader to assess the magnitude of this statistic without the time-consuming and cumbersome process of looking up the critical values from a table. Instead, if you write that the p-value is $p_n = 0.19$ (as is the case for $W_n = 11.2$ and $q = 8$) then it is simple for a reader to intrepret its magnitude as "insignificant".

For example, 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 (8.10) for this joint hypothesis and find $W_n = 23.14$ with a p-value of $p_n = 0.000$. Thus we reject the 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 the joint Wald statistic and the individual t-statistics is useful for interpretation.

Abraham Wald

The Hungarian mathematician/statistician/econometrician Abraham Wald (1902-1950) developed an optimality property forl the Wald test in terms of weighted average power. He also developed the field of sequential testing and the design of experiments.

8.10 Homoskedastic Wald Tests

If the error is known to be homoskedastic, then it is appropriate to use the homoskedastic Wald statistic (6.49) which replaces $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\theta}}}$ with the homoskedastic estimate $\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\theta}}}$. This statistic equals

$$\begin{aligned} W_n^0 &= \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{V}}^0_{\widehat{\boldsymbol{\theta}}}\right)^{-1} \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\ &= \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{R}}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \widehat{\boldsymbol{R}}\right)^{-1} \left(\boldsymbol{r}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\theta}_0\right) / s^2. \qquad (8.12) \end{aligned}$$

In the case of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$ we can write this as

$$W_n^0 = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)' \left(\widehat{\boldsymbol{R}}' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right) / s^2. \qquad (8.13)$$

We call (8.12) or (8.13) a **homoskedastic Wald statistic** as it is an appropriate test when the errors are conditionally homoskedastic.

As for W_n, when $q = 1$ then $W_n^0 = t_n^2$, the square of the t-statistic where the latter is computed with a homoskedastic standard error.

Theorem 8.10.1 *Under Assumptions 6.1.2 and 6.10.1,* $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, *and* $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, *then*

$$W_n^0 \xrightarrow{d} \chi_q^2,$$

and for c satisfying $\alpha = 1 - G_q(c)$,

$$\Pr\left(W_n^0 > c \mid \mathbb{H}_0\right) \longrightarrow \alpha$$

so the test "Reject $\mathbb{H}_0$ *if* $W_n^0 > c$*" has asymptotic size* α.

8.11 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 $J_n(\boldsymbol{\beta})$ 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 $\mathbb{H}_0:\boldsymbol{\beta}\notin\boldsymbol{B}_0$ 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

$$\begin{aligned}\widehat{\boldsymbol{\beta}}&=\underset{\boldsymbol{\beta}\in\boldsymbol{B}}{\operatorname{argmin}}\ J_n(\boldsymbol{\beta})\\ \widetilde{\boldsymbol{\beta}}&=\underset{\boldsymbol{\beta}\in\boldsymbol{B}_0}{\operatorname{argmin}}\ J_n(\boldsymbol{\beta}).\end{aligned}$$

The **criterion-based statistic** for $\mathbb{H}_0$ versus $\mathbb{H}_1$ is proportional to

$$\begin{aligned}J_n&=\min_{\boldsymbol{\beta}\in\boldsymbol{B}_0}J_n(\boldsymbol{\beta})-\min_{\boldsymbol{\beta}\in\boldsymbol{B}}J_n(\boldsymbol{\beta})\\ &=J_n(\widetilde{\boldsymbol{\beta}})-J_n(\widehat{\boldsymbol{\beta}}).\end{aligned}$$

The criterion-based statistic J_n is sometimes called a **distance** statistic, a **minimum-distance** statistic, or a **likelihood-ratio-like** statistic.

Since $\boldsymbol{B}_0$ is a subset of $\boldsymbol{B}$, $J_n(\widetilde{\boldsymbol{\beta}})\geq J_n(\widehat{\boldsymbol{\beta}})$ and thus $J_n\geq 0$. The statistic J_n measures the cost (on the criterion) of imposing the null restriction $\boldsymbol{\beta}\in\boldsymbol{B}_0$.

8.12 Minimum Distance Tests

The minimum distance test is a criterion-based test where $J_n(\boldsymbol{\beta})$ is the minimum distance criterion (7.17)

$$J_n(\boldsymbol{\beta})=n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\boldsymbol{W}_n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\tag{8.14}$$

with $\widehat{\boldsymbol{\beta}}$ the unrestricted (LS) estimator. The restricted estimator $\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}$ minimizes (8.14) subject to $\boldsymbol{\beta}\in\boldsymbol{B}_0$. Observing that $J_n(\widehat{\boldsymbol{\beta}})=0$, the minimum distance statistic simplifies to

$$J_n=J_n(\widetilde{\boldsymbol{\beta}}_{\mathrm{md}})=n\left(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}\right)'\boldsymbol{W}_n\left(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\mathrm{md}}\right).\tag{8.15}$$

The efficient minimum distance estimator $\widetilde{\boldsymbol{\beta}}_{\mathrm{emd}}$ is obtained by setting $\boldsymbol{W}_n=\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}$ in (8.14) and (8.15). The efficient minimum distance statistic for $\mathbb{H}_0:\boldsymbol{\beta}\in\boldsymbol{B}_0$ is therefore

$$J_n^*=n\left(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\mathrm{emd}}\right)'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\left(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\mathrm{emd}}\right).\tag{8.16}$$

Consider the class of linear hypotheses $\mathbb{H}_0:\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{\theta}_0$. In this case we know from (7.25) that the efficient minimum distance estimator $\widetilde{\boldsymbol{\beta}}_{\mathrm{emd}}$ subject to the constraint $\boldsymbol{R}'\boldsymbol{\beta}=\boldsymbol{\theta}_0$ is

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{emd}}=\widehat{\boldsymbol{\beta}}-\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}-\boldsymbol{\theta}_0\right)$$

and thus

$$\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\mathrm{emd}}=\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}-\boldsymbol{\theta}_0\right).$$

Substituting into (8.16) we find

$$\begin{aligned} J_n^* &= n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}-\boldsymbol{\theta}_0\right)'\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}-\boldsymbol{\theta}_0\right) \\ &= n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}-\boldsymbol{\theta}_0\right)'\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}}-\boldsymbol{\theta}_0\right) \\ &= W_n, \end{aligned} \tag{8.17}$$

which is the Wald statistic (8.10).

Thus for linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$, the efficient minimum distance statistic J_n^* is identical to the Wald statistic (8.10). For non-linear hypotheses, however, the Wald and minimum distance statistics are different.

Newey and West (1987) established the asymptotic null distribution of J_n^* for linear and non-linear hypotheses.

Theorem 8.12.1 *Under Assumptions 6.1.2 and 6.10.1, and* $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, *then* $J_n^* \xrightarrow{d} \chi_q^2$.

Testing using the minimum distance statistic J_n^* is similar to testing using the Wald statistic W_n. Critical values and p-values are computed using the χ_q^2 distribution. $\mathbb{H}_0$ is rejected in favor of $\mathbb{H}_1$ if J_n^* exceeds the level α critical value. The asymptotic p-value is $p_n = 1 - F_q(J_n^*)$.

8.13 Minimum Distance Tests Under Homoskedasticity

If we set $\boldsymbol{W}_n = \widehat{\boldsymbol{Q}}_{xx}$ in (8.14) we obtain the criterion (7.19)

$$J_n^0(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\widehat{\boldsymbol{Q}}_{xx}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right).$$

A minimum distance statistic for $\mathbb{H}_0 : \boldsymbol{\beta} \in \boldsymbol{B}_0$ is

$$J_n^0 = \min_{\boldsymbol{\beta}\in\boldsymbol{B}_0} J_n^0(\boldsymbol{\beta})/s^2.$$

Notice that we have scaled the criterion by the unbiased variance estimator s^2 from (4.21) for reasons which will become clear momentarily.

Equation (7.20) showed that

$$SSE_n(\boldsymbol{\beta}) = n\hat{\sigma}^2 + J_n^0(\boldsymbol{\beta})$$

and so the minimizers of $SSE_n(\boldsymbol{\beta})$ and $J_n^0(\boldsymbol{\beta})$ are identical. Thus the constrained minimizer of $J_n^0(\boldsymbol{\beta})$ is constrained least-squares

$$\widetilde{\boldsymbol{\beta}}_{\text{cls}} = \underset{\boldsymbol{\beta}\in\boldsymbol{B}_0}{\operatorname{argmin}}\, J_n^0(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\boldsymbol{B}_0}{\operatorname{argmin}}\, SSE_n(\boldsymbol{\beta}) \tag{8.18}$$

and therefore

$$\begin{aligned} J_n^0 =&\ J_n^0(\widetilde{\boldsymbol{\beta}}_{\text{cls}})/s^2 \\ &= n\left(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)'\widehat{\boldsymbol{Q}}_{xx}\left(\widehat{\boldsymbol{\beta}}-\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right)/s^2. \end{aligned}$$

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 (7.10)

$$\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)$$

and solving we find

$$J_n^0 = n\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)'\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{xx}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{\theta}_0\right)/s^2 = W_n^0. \tag{8.19}$$

This is the homoskedastic Wald statistic (8.13). 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 8.13.1 *Under Assumptions 6.1.2 and 6.10.1,* $\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right) = \sigma^2$, *and* $\mathbb{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, *then* $J_n^0 \xrightarrow{d} \chi_q^2$.

8.14 F Tests

The F statistic for testing $\mathbb{H}_0 : \boldsymbol{\beta} \in \boldsymbol{B}_0$ is

$$F_n = \frac{\left(SSE_n(\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}}) - SSE_n(\widehat{\boldsymbol{\beta}})\right)/q}{SSE_n(\widehat{\boldsymbol{\beta}})/(n-k)} \tag{8.20}$$

where

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

is the sum-of-squared errors, $\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}}$ is the constrained least-squares estimator (8.18), $\widehat{\boldsymbol{\beta}}$ is the unconstrained least-squares estimator, q is the number of restrictions, and k is the number of unconstrained coefficients.

Noting that $s^2 = SSE_n(\widehat{\boldsymbol{\beta}})/(n-k)$, we can also write (8.20) as

$$F_n = \frac{SSE_n(\widetilde{\boldsymbol{\beta}}_{\mathrm{cls}}) - SSE_n(\widehat{\boldsymbol{\beta}})}{qs^2}$$

which is a scale of the difference of sum-of-squared errors, and is thus a criterion-based statistic. Using (7.20) we can also write the statistic as

$$F_n = J_n^0/q,$$

so the F stastistic is identical to the homoskedastic minimum distance statistic divided by the number of restrictions q.

Another useful way of writing (8.20) is

$$F_n = \left(\frac{n-k}{q}\right)\frac{\left(\tilde{\sigma}^2 - \hat{\sigma}^2\right)}{\hat{\sigma}^2} \tag{8.21}$$

where

$$\hat{\sigma}^2 = \frac{SSE_n(\widehat{\boldsymbol{\beta}})}{n} = \frac{1}{n}\sum_{i=1}^n \hat{e}_i^2$$

is the residual variance estimate under $\mathbb{H}_1$ and

$$\tilde{\sigma}^2 = \frac{SSE_n(\widetilde{\boldsymbol{\beta}}_{\text{cls}})}{n} = \frac{1}{n}\sum_{i=1}^{n}\tilde{e}_i^2$$

with $\tilde{e}_i = y_i - \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}}_{\text{cls}}$ is the residual variance estimate under $\mathbb{H}_0$.

As we discussed in the previous section, in the special case of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\boldsymbol{\beta} = \boldsymbol{\theta}_0$, $J_n^0 = W_n^0$. It follows that in this case $F_n = W_n^0/q$. Thus the F_n statistic equals the homoskedastic Wald statistic divided by q. It follows that they are equivalent tests for $\mathbb{H}_0$ against $\mathbb{H}_1$.

In many statistical packages, linear hypothesis tests are reported as F statistics rather than Wald statistics. While they are equivalent, it is important to know which is being reported to know which critical values to use. (If p-values are directly reported this is not an issue.)

When reporting an F_n statistic it is conventional to calculate critical values and p-values using the $F(q, n-k)$ distribution instead of the asymptotic χ_q^2/q distribution. This is a prudent small sample adjustment, as the F distribution is exact when the errors are independent of the regressors and normally distributed. However, when the degrees of freedom $n-k$ are large then the difference is negligible. More relevantly, if $n-k$ is small enough to make a difference, probably we shouldn't be trusting the asymptotic approximation anyway!

An elegant feature about (8.20) or (8.21) 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_n 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_n 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 $q \geq 7$.

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.

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.

[1] All coefficients except the intercept.

8.15 Likelihood Ratio Test

For a model with parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and likelihood function $L_n(\boldsymbol{\theta})$ the likelihood ratio statistic for $\mathbb{H}_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ versus $\mathbb{H}_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$ is

$$\begin{aligned} LR_n &= 2\left(\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \log L_n(\boldsymbol{\theta}) - \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \log L_n(\boldsymbol{\theta})\right) \\ &= 2\left(\log L_n(\widehat{\boldsymbol{\theta}}) - \log L_n(\widetilde{\boldsymbol{\theta}})\right) \end{aligned}$$

where $\widehat{\boldsymbol{\theta}}$ and $\widetilde{\boldsymbol{\theta}}$ are the unrestricted and constrained MLE.

In the normal linear model the maximized log likelihood (3.45) at the unrestricted and restricted estimates are

$$\log L\left(\widehat{\boldsymbol{\beta}}, \hat{\sigma}^2\right) = -\frac{n}{2}\left(\log(2\pi) + 1\right) - \frac{n}{2}\log\left(\hat{\sigma}^2\right)$$

and

$$\log L\left(\widetilde{\boldsymbol{\beta}}, \tilde{\sigma}^2\right) = -\frac{n}{2}\left(\log(2\pi) + 1\right) - \frac{n}{2}\log\left(\tilde{\sigma}^2\right)$$

respectively. Thus the LR statistic is

$$\begin{aligned} LR_n &= n\left(\log\left(\tilde{\sigma}^2\right) - \log\left(\hat{\sigma}^2\right)\right) \\ &= n \log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right) \end{aligned}$$

which is a monotonic function of $\tilde{\sigma}^2/\hat{\sigma}^2$. Recall that the F statistic (8.21) is also a monotonic function of $\tilde{\sigma}^2/\hat{\sigma}^2$. Thus LR_n and F_n are fundamentally the same statistic and have the same information about $\mathbb{H}_0$ versus $\mathbb{H}_1$.

Furthermore, by a first-order Taylor series approximation

$$LR_n/q = \frac{n}{q}\log\left(1 + \frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1\right) \simeq \frac{n}{q}\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} - 1\right) \simeq F_n.$$

This shows that the two statistics (LR_n and F_n) will be numerically close. It also shows that the F statistic and the homoskedastic Wald statistic for linear hypotheses can also be interpreted as approximate likelihood ratio statistics under normality.

8.16 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 $\hat{\beta}$ and $\hat{\sigma}^2$ be the sample mean and variance of y_i. The standard Wald test for $\mathbb{H}_0$ is

$$W_n = n\frac{\left(\hat{\beta} - 1\right)^2}{\hat{\sigma}^2}.$$

Now notice that $\mathbb{H}_0$ is equivalent to the hypothesis

$$\mathbb{H}_0(s) : \beta^s = 1$$

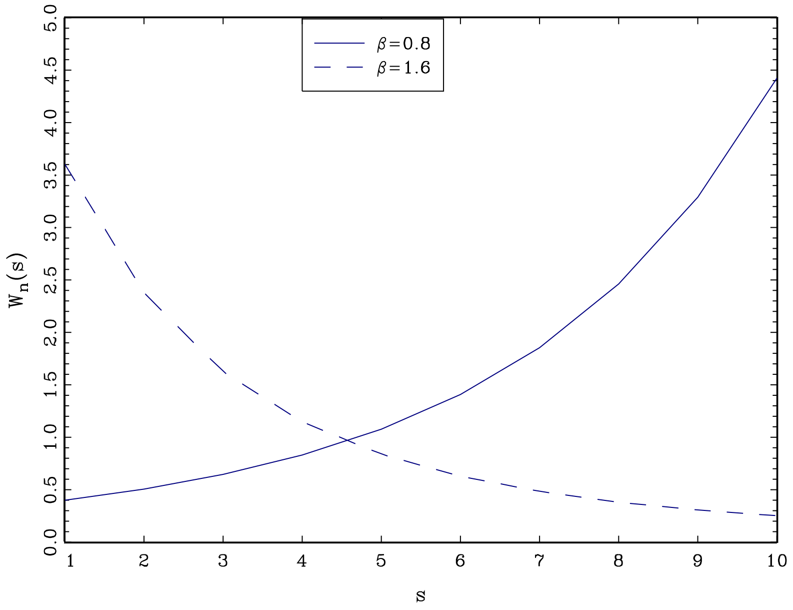


Figure 8.1: Wald Statistic as a function of s

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

$$W_n(s) = n\frac{\left(\hat{\beta}^s - 1\right)^2}{\hat{\sigma}^2 s^2 \hat{\beta}^{2s-2}}.$$

While the hypothesis $\beta^s = 1$ is unaffected by the choice of s, the statistic $W_n(s)$ varies with s. This is an unfortunate feature of the Wald statistic.

To demonstrate this effect, we have plotted in Figure 8.1 the Wald statistic $W_n(s)$ as a function of s, setting $n/\hat{\sigma}^2 = 10$. The increasing solid line is for the case $\hat{\beta} = 0.8$. The decreasing dashed line is for the case $\hat{\beta} = 1.6$. 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 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 $W_n(s) \xrightarrow{d} \chi_1^2$ 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, $\Pr\left(W_n(s) > 3.84 \mid \beta = 1\right)$. Given the simplicity of the model, this probability depends only on s, n, and σ^2. In Table 8.1 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 $W_n(s)$ 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 $W_n = W_n(1)$ test. Any other choice of s leads to a test with unacceptable Type I error probabilities.

Table 8.1

Type I Error Probability of Asymptotic 5% $W_n(s)$ 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 8.1 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 $W_n(s)$ on s as shown in Figure 8.1.

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

$$\begin{aligned} y_i &= \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= \mathbf{0} \end{aligned} \tag{8.22}$$

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 $\widehat{\boldsymbol{\beta}} = (\hat\beta_0, \hat\beta_1, \hat\beta_2)$ be the least-squares estimates of (8.22), let $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ be an estimate of the covariance matrix for $\widehat{\boldsymbol{\beta}}$ and set $\hat\theta = \hat\beta_1/\hat\beta_2$. Define

$$\widehat{\boldsymbol{R}}_1 = \begin{pmatrix} 0 \\ \\ \dfrac{1}{\hat\beta_2} \\ \\ -\dfrac{\hat\beta_1}{\hat\beta_2^2} \end{pmatrix}$$

so that the standard error for $\hat{\theta}$ is $s(\hat{\theta}) = \left(\widehat{\boldsymbol{R}}_1' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}_1\right)^{1/2}$. In this case a t-statistic for $\mathbb{H}_0$ is

$$t_{1n} = \frac{\left(\frac{\hat{\beta}_1}{\hat{\beta}_2} - \theta_0\right)}{s(\hat{\theta})}.$$

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

$$t_{2n} = \frac{\hat{\beta}_1 - \theta_0 \hat{\beta}_2}{\left(\boldsymbol{R}_2' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \boldsymbol{R}_2\right)^{1/2}}.$$

where

$$\boldsymbol{R}_2 = \begin{pmatrix} 0 \\ 1 \\ -\theta_0 \end{pmatrix}.$$

To compare t_{1n} and t_{2n} 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 8.2
Type I Error Probability of Asymptotic 5% t-tests

	$n = 100$				$n = 500$			
	$\Pr(t_n < -1.645)$		$\Pr(t_n > 1.645)$		$\Pr(t_n < -1.645)$		$\Pr(t_n > 1.645)$	
β_2	t_{1n}	t_{2n}	t_{1n}	t_{2n}	t_{1n}	t_{2n}	t_{1n}	t_{2n}
.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 $\Pr(t_n < -1.645)$ and $\Pr(t_n > 1.645)$ are calculated from 50,000 simulated samples. The results are presented in Table 8.2. Ideally, the entries in the table should be 0.05. However, the rejection rates for the t_{1n} statistic diverge greatly from this value, especially for small values of β_2. The left tail probabilities $\Pr(t_{1n} < -1.645)$ greatly exceed 5%, while the right tail probabilities $\Pr(t_{1n} > 1.645)$ are close to zero in most cases. In contrast, the rejection rates for the linear t_{2n} 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 4.2 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_n, which equals W_n with $r = 1$ in the first example, and $|t_{2n}|$ 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.

8.17 Monte Carlo Simulation

In the Section 8.16 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 $T_n = T_n\left((y_1, \boldsymbol{x}_1), ..., (y_n, \boldsymbol{x}_n), \boldsymbol{\theta}\right)$ be a statistic of interest, for example an estimator $\hat{\theta}$ or a t-statistic $(\hat{\theta} - \theta)/s(\hat{\theta})$. The exact distribution of T_n is

$$G_n(u, F) = \Pr\left(T_n \leq u \mid F\right).$$

While the asymptotic distribution of T_n might be known, the exact (finite sample) distribution G_n is generally unknown.

Monte Carlo simulation uses numerical simulation to compute $G_n(u, F)$ for selected choices of F. This is useful to investigate the performance of the statistic T_n in reasonable situations and sample sizes. The basic idea is that for any given F, the distribution function $G_n(u, F)$ 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 $T_n = T_n\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), \boldsymbol{\theta}\right)$ is calculated on this pseudo data.

For step 1, most computer packages have built-in procedures for generating U$[0, 1]$ and N$(0, 1)$ random numbers, and 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 $G_n(u, F)$. 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_{nb}, $b = 1, ..., B$. These results are stored. After all B experiments have been calculated, these results constitute a random sample of size B from the distribution of $G_n(u, F) = \Pr\left(T_{nb} \leq u\right) = \Pr\left(T_n \leq u \mid F\right)$.

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 $\hat{\theta} - \theta$. We then set $T_n = \hat{\theta} - \theta$, run the above experiment, and calculate

$$\widehat{Bias(\hat{\theta})} = \frac{1}{B}\sum_{b=1}^{B} T_{nb} = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b - \theta$$

$$\widehat{MSE(\hat{\theta})} = \frac{1}{B}\sum_{b=1}^{B} (T_{nb})^2 = \frac{1}{B}\sum_{b=1}^{B} \left(\hat{\theta}_b - \theta\right)^2$$

$$\widehat{\operatorname{var}(\hat{\theta})} = \widehat{MSE(\hat{\theta})} - \left(\widehat{Bias(\hat{\theta})}\right)^2$$

Suppose we are interested in the Type I error associated with an asymptotic 5% two-sided t-test. We would then set $T_n = \left|\hat{\theta} - \theta\right| / s(\hat{\theta})$ and calculate

$$\hat{P} = \frac{1}{B}\sum_{b=1}^{B} 1\left(T_{nb} \geq 1.96\right), \tag{8.23}$$

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 $T_n = \hat{\theta}$ or $T_n = \left(\hat{\theta} - \theta\right) / s(\hat{\theta})$ We then compute the 5% and 95% sample quantiles of the sample $\{T_{nb}\}$. The $\alpha\%$ sample quantile is a number q_α such that $\alpha\%$ of the sample are less than q_α. A simple way to compute sample quantiles is to sort the sample $\{T_{nb}\}$ 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_n, 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 (8.23). The random variable $1\left(T_{nb} \geq 1.96\right)$ is iid Bernoulli, equalling 1 with probability $p = \mathbb{E}1\left(T_{nb} \geq 1.96\right)$. The average (8.23) is therefore an unbiased estimator of p with standard error $s\left(\hat{p}\right) = \sqrt{p\left(1-p\right)/B}$. As p is unknown, this may be approximated by replacing p with $\hat{p}$ or with an hypothesized value. For example, if we are assessing an asymptotic 5% test, then we can set $s\left(\hat{p}\right) = \sqrt{(.05)\left(.95\right)/B} \simeq .22/\sqrt{B}$. Hence, standard errors for $B = 100$, 1000, and 5000, are, respectively, $s\left(\hat{p}\right) = .022$, .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, $\hat{p}_1 = 0.07$ and $\hat{p}_2 = 0.03$, of a common probability p. The asymptotic (as $B \to \infty$) distribution of their difference is $\sqrt{B}\left(\hat{p}_1 - \hat{p}_2\right) \xrightarrow{d} \mathrm{N}(0, 2p(1-p))$, so a standard error for $\hat{p}_1 - \hat{p}_2 = 0.04$ is $\hat{s} = \sqrt{2p(1-p)/B} \simeq 0.03$, using the estimate $p = (\hat{p}_1 + \hat{p}_2)/2$. 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 $\hat{p}_1 = 0.07$ and $\hat{p}_2 = 0.03$ 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 $\hat{p}_2 = 0.04$. Then $\hat{p}_1 - \hat{p}_2 = 0.03$ and a standard error is $\hat{s} = \sqrt{p(1-p)\left(1/100 + 1/5000\right)} \simeq 0.02$. 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. Never-the-less, 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.

8.18 Confidence Intervals by Test Inversion

There is a close relationship between hypothesis tests and confidence intervals. We observed in Section 6.13 that the standard 95% asymptotic confidence interval for a parameter θ is

$$\begin{aligned} C_n &= \left[\widehat{\theta} - 1.96 \cdot s(\widehat{\theta}), \quad \widehat{\theta} + 1.96 \cdot s(\widehat{\theta})\right] \\ &= \{\theta : |t_n(\theta)| \leq 1.96\}. \end{aligned} \tag{8.24}$$

That is, we can describe C_n 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 $T_n(\theta)$ and critical value c, the acceptance region "Accept if $T_n(\theta) \leq c$" is identical to the confidence interval $C_n = \{\theta : T_n(\theta) \leq c\}$. Since the regions are identical, the probability of coverage $\Pr(\theta \in C_n)$ 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 C_n as in (8.24) where $\hat{\theta} = r(\widehat{\boldsymbol{\beta}})$ is the point estimate and $s(\hat{\theta}) = \sqrt{\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \widehat{\boldsymbol{R}}}$ 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 8.16 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 (8.24) 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 8.16 – 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

$$t_n(\theta) = \frac{\hat{\beta}_1 - \hat{\beta}_2\theta}{\left(\boldsymbol{R}' \widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}} \boldsymbol{R}\right)^{1/2}}$$

where

$$\boldsymbol{R} = \begin{pmatrix} 1 \\ -\theta \end{pmatrix}$$

and $\widehat{\boldsymbol{V}}_{\widehat{\boldsymbol{\beta}}}$ is the covariance matrix for $(\hat{\beta}_1\ \hat{\beta}_2)$. A 95% confidence interval for $\theta = \beta_1/\beta_2$ is the set of values of θ such that $|t_n(\theta)| \leq 1.96$. Since θ appears in both the numerator and denominator, $t_n(\theta)$ 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 $\hat{\theta} = 29.8$ and we can calculate the standard error $s(\hat{\theta}) = 0.022$ 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 8.16 we know the first interval can be quite inaccurate and the second interval is greatly preferred.

8.19 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 $t_n(\theta) = \sqrt{n}\left(\bar{y} - \theta\right)/\sigma$, and tests of $\mathbb{H}_0 : \theta = 0$ against $\mathbb{H}_1 : \theta > 0$. We reject $\mathbb{H}_0$ if $t_n = t_n(0) > c$. Note that

$$t_n = t_n(\theta) + \sqrt{n}\theta/\sigma$$

and $t_n(\theta)$ has an exact $\mathrm{N}(0, 1)$ distribution. This is because $t_n(\theta)$ is centered at the true mean θ, while the test statistic $t_n(0)$ is centered at the (false) hypothesized mean of 0.

The power of the test is

$$\Pr\left(t_n > c \mid \theta\right) = \Pr\left(\mathrm{Z} + \sqrt{n}\theta/\sigma > c\right) = 1 - \Phi\left(c - \sqrt{n}\theta/\sigma\right).$$

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 8.19.1 *A test of* $\mathbb{H}_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ *is **consistent against fixed alternatives** if for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, $\Pr\left(\textit{Reject } \mathbb{H}_0 \mid \theta\right) \to 1$ *as* $n \to \infty$.

For tests of the form "Reject $\mathbb{H}_0$ if $T_n > c$", a sufficient condition for test consistency is that the T_n diverges to positive infinity with probability one for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_1$.

Definition 8.19.2 $T_n \xrightarrow{p} \infty$ *as* $n \to \infty$ *if for all* $M < \infty$, $\Pr\left(T_n \leq M\right) \to 0$ *as* $n \to \infty$. *Similarly,* $T_n \xrightarrow{p} -\infty$ *as* $n \to \infty$ *if for all* $M < \infty$, $\Pr\left(T_n \geq -M\right) \to 0$ *as* $n \to \infty$.

In general, t-tests and Wald tests are consistent against fixed alternatives. Take a t-statistic for a test of $\mathbb{H}_0 : \theta = \theta_0$

$$t_n = \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})}$$

where θ_0 is a known value and $s(\widehat{\theta}) = \sqrt{n^{-1}\hat{V}_\theta}$. Note that

$$t_n = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} + \frac{\sqrt{n}\left(\theta - \theta_0\right)}{\sqrt{\hat{V}_\theta}}.$$

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 8.19.1 *Under Assumptions 6.1.2 and 6.10.1, for* $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$ *and* $q = 1$*, then* $|t_n| \xrightarrow{p} \infty$*, so for any* $c < \infty$ *the test "Reject* $\mathbb{H}_0$ *if* $|t_n| > c$*" 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

$$W_n = n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right).$$

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 $\mathbb{H}_1$, $W_n \xrightarrow{p} \infty$. Again, this implies that Wald tests are consistent tests.

Theorem 8.19.2 *Under Assumptions 6.1.2 and 6.10.1, for* $\boldsymbol{\theta} = \boldsymbol{r}(\boldsymbol{\beta}) \neq \boldsymbol{\theta}_0$*, then* $W_n \xrightarrow{p} \infty$*, so for any* $c < \infty$ *the test "Reject* $\mathbb{H}_0$ *if* $W_n > c$*" is consistent against fixed alternatives.*

8.20 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 7.9). 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

$$\theta_n = r(\boldsymbol{\beta}_n) = \theta_0 + n^{-1/2}h \tag{8.25}$$

where the scalar h is 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 (8.25) 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 (8.25) 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 (8.25) might seem odd at first as in the actual world the sample size cannot mechanically affect the value of the parameter. Thus (8.25) 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 (8.25). 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

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\xrightarrow{d}\sqrt{V_\theta}\mathrm{Z}+h$$

or $\mathrm{N}(h,V_\theta)$. This is a continuous asymptotic distribution, and depends continuously on the localing parameter h.

Applied to the t statistic we find

$$\begin{aligned}t_n &= \frac{\widehat{\theta}-\theta_0}{s(\widehat{\theta})}\\ &\xrightarrow{d}\frac{\sqrt{V_\theta}\mathrm{Z}+h}{\sqrt{V_\theta}}\\ &\sim \mathrm{Z}+\delta \qquad (8.26)\end{aligned}$$

where $\delta = h/\sqrt{V_\theta}$. This generalizes Theorem 8.4.1 (which assumes $\mathbb{H}_0$ is true) to allow for local alternatives of the form (8.25).

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 $t_n>c_\alpha$ where $\Phi(c_\alpha)=1-\alpha$. The **asymptotic local power** of this test is the limit (as the sample size diverges) of the rejection probability under the local alternative (8.25)

$$\begin{aligned}\lim_{n\to\infty}\Pr\left(\text{Reject }\mathbb{H}_0\right) &= \lim_{n\to\infty}\Pr\left(t_n>c_\alpha\right)\\ &= \Pr\left(\mathrm{Z}+\delta>c_\alpha\right)\\ &= 1-\Phi\left(c_\alpha-\delta\right)\\ &= \Phi\left(\delta-c_\alpha\right)\\ &\overset{def}{=}\pi_\alpha(\delta).\end{aligned}$$

We call $\pi_\alpha(\delta)$ the **local power function**.

In Figure 8.2 we plot the local power function $\pi_\alpha(\delta)$ 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 $\pi_\alpha(\delta)=\alpha$. 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 $s(\hat{\theta})$. To see this, recall that $s(\hat{\theta})=n^{-1/2}\sqrt{\hat{V}_\theta}\simeq n^{-1/2}\sqrt{V_\theta}$ and then note that

$$\delta=\frac{h}{\sqrt{V_\theta}}\simeq\frac{n^{-1/2}h}{s(\hat{\theta})}=\frac{\theta_n-\theta_0}{s(\hat{\theta})}.$$

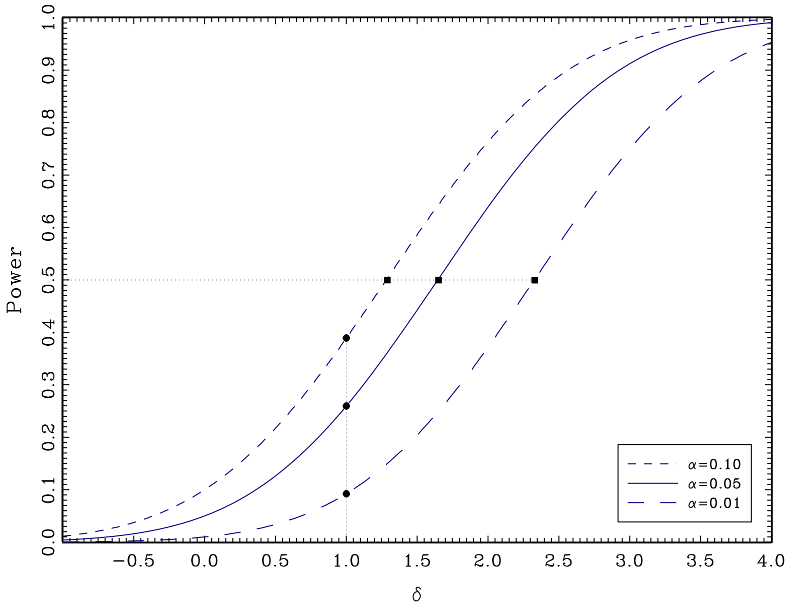


Figure 8.2: Asymptotic Local Power Function of One-Sided t Test

Thus δ approximately equals the deviation $\theta_n - \theta_0$ expressed as multiples of the standard error $s(\hat{\theta})$. Thus as we examine Figure 8.2, 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 0.008. Thus in this example, $\delta = 1$ corresonds to $\theta_n = 0.008$ or an 0.8% 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 8.2 there is a vertical dotted line at $\delta = 1$, showing that the asymptotic local power function $\pi_\alpha(\delta)$ 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 8.2 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 (8.25) $\delta = h/\sqrt{V_\theta} = \sqrt{n}\,(\theta_n - \theta_0)/\sqrt{V_\theta}$. Thus holding θ and V_θ fixed, δ^2 is proportional to n.

The analysis of a two-sided t test is similar. (8.26) implies that

$$t_n = \left| \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})} \right| \xrightarrow{d} |\mathrm{Z} + \delta|$$

and thus the local power of a two-sided t test is

$$\begin{aligned}\lim_{n\to\infty} \Pr(\text{Reject } \mathbb{H}_0) &= \lim_{n\to\infty} \Pr(t_n > c_\alpha) \\ &= \Pr(|\mathrm{Z}+\delta| > c_\alpha) \\ &= \Phi(\delta - c_\alpha) - \Phi(-\delta - c_\alpha)\end{aligned}$$

which is monotonically increasing in $|\delta|$.

Theorem 8.20.1 *Under Assumptions 6.1.2 and 6.10.1, 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_α *such that* $\Pr(Z > c_\alpha) = \alpha$,

$$\Pr(t(\theta_0) > c_\alpha) \longrightarrow \Phi(\delta - c_\alpha).$$

Furthermore, for c_α *such that* $\Pr(|\mathrm{Z}| > c_\alpha) = \alpha$,

$$\Pr(|t(\theta_0)| > c_\alpha) \longrightarrow \Phi(\delta - c_\alpha) - \Phi(-\delta - c_\alpha).$$

8.21 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 (8.25) to allow $\boldsymbol{\theta}_n$ to be vector-valued. The local parameterization takes the form

$$\boldsymbol{\theta}_n = \boldsymbol{r}(\boldsymbol{\beta}_n) = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h} \tag{8.27}$$

where $\boldsymbol{h}$ is $q \times 1$.

Under (8.27),

$$\begin{aligned}\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= \sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_n\right) + \boldsymbol{h} \\ &\xrightarrow{d} \mathrm{Z}_{\boldsymbol{h}} \sim \mathrm{N}(\boldsymbol{h}, \boldsymbol{V}_{\boldsymbol{\theta}}),\end{aligned}$$

a normal random vector with mean $\boldsymbol{h}$ and variance matrix $\boldsymbol{V}_{\boldsymbol{\theta}}$.

Applied to the Wald statistic we find

$$\begin{aligned}W_n &= n\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \widehat{\boldsymbol{V}}_{\boldsymbol{\theta}}^{-1}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \\ &\xrightarrow{d} \mathrm{Z}_{\boldsymbol{h}}' \boldsymbol{V}_{\boldsymbol{\theta}}^{-1} \mathrm{Z}_{\boldsymbol{h}}.\end{aligned} \tag{8.28}$$

The asymptotic distribution (8.28) is a quadratic in the normal random vector $\mathrm{Z}_{\boldsymbol{h}}$, similar to that found for the asymptotic null distribution of the Wald statistic. The important difference, however, is that $\mathrm{Z}_{\boldsymbol{h}}$ has a mean of $\boldsymbol{h}$. The distribution of the quadratic form (8.28) is a close relative of the chi-square distribution.

Theorem 8.21.1 *If* $\mathrm{Z}_{\boldsymbol{h}} \sim \mathrm{N}(\boldsymbol{h}, \boldsymbol{V})$ *with* $\boldsymbol{V} > 0$, $q \times q$, *then* $\mathrm{Z}_{\boldsymbol{h}}' \boldsymbol{V}^{-1} \mathrm{Z}_{\boldsymbol{h}} \sim \chi_q^2(\lambda)$, *a n**on-central chi-square random variable** with q degrees of freedom and non-centrality parameter* $\lambda = \boldsymbol{h}' \boldsymbol{V}^{-1} \boldsymbol{h}$.

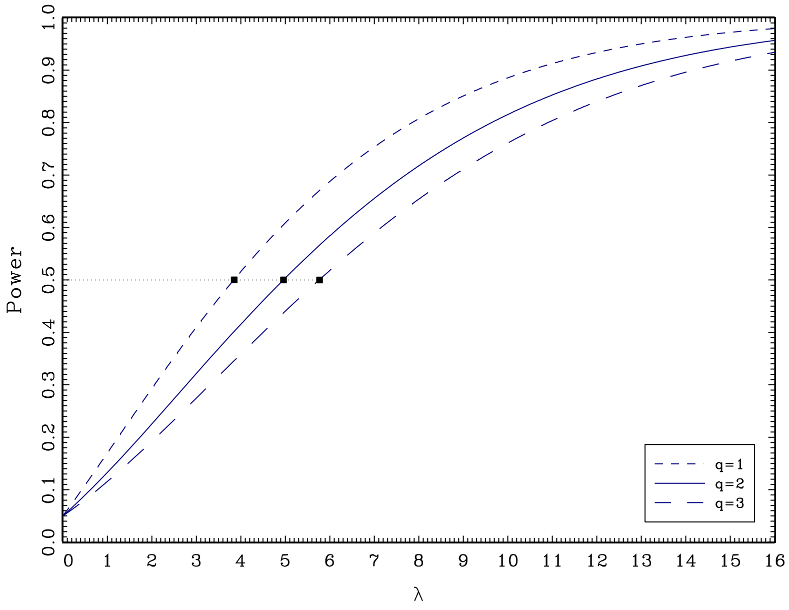


Figure 8.3: Asymptotic Local Power Function, Varying q

The convergence (8.28) shows that under the local alternatives (8.27), $W_n \xrightarrow{d} \chi_q^2(\lambda)$. 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_α so that $\Pr\left(\chi_q^2 > c_\alpha\right) = \alpha$. Then as $n \to \infty$,

$$\Pr\left(W_n > c_\alpha\right) \longrightarrow \Pr\left(\chi_q^2(\lambda) > c_\alpha\right) \overset{def}{=} \pi_{\alpha,q}(\lambda).$$

The asymptotic local power function $\pi_{\alpha,q}(\lambda)$ depends only on α, q, and λ.

Theorem 8.21.2 *Under Assumptions 6.1.2 and 6.10.1, and* $\boldsymbol{\theta}_n = \boldsymbol{r}(\boldsymbol{\beta}_n) = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{h}$, *then*

$$W_n \xrightarrow{d} \chi_q^2(\lambda)$$

where $\lambda = \boldsymbol{h}'\boldsymbol{V}_{\boldsymbol{\theta}}^{-1}\boldsymbol{h}$. *Furthermore, for* c_α *such that* $\Pr\left(\chi_q^2 > c_\alpha\right) = \alpha$,

$$\Pr\left(W_n > c_\alpha\right) \longrightarrow \Pr\left(\chi_q^2(\lambda) > c_\alpha\right).$$

The non-central chi-square distribution is a generalization of the chi-square, with $\chi_q^2(\lambda)$ specializing to χ_q^2 when $\lambda = 0$. In the case $q = 1$, $\chi_q^2(\lambda) = |\mathrm{Z} + \delta|^2$ with $\lambda = \delta^2$, and thus Theorem 8.21.2 generalizes Theorem 8.20.1 from $q = 1$ to $q \geq 1$.

Figure 8.3 plots $\pi_{0.05,q}(\lambda)$ (the power of asymptotic 5% tests) as a function of λ for $q = 1$, $q = 2$, and $q = 3$. The power functions are monotonically increasing in λ and asymptote to one.

Figure 8.3 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.

8.22 Technical Proofs*

Proof of Theorem 8.12. The conditions of Theorem 7.10.1 hold, since $\mathbb{H}_0$ implies Assumption 7.5.1. From (7.54) with $\boldsymbol{W}_n^{-1} = \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}$, 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 $\mathrm{Z} \sim \mathrm{N}(\boldsymbol{0}, (\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R})^{-1})$. Thus

$$\begin{aligned}J_n^* &= n\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right)'\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\left(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_{\text{emd}}\right) \\ &\xrightarrow{d} \mathrm{Z}'\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\,\mathrm{Z} \\ &= \mathrm{Z}'\left(\boldsymbol{R}'\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{R}\right)\,\mathrm{Z} \\ &= \chi_q^2.\end{aligned}$$

■

Proof of Theorem 8.21.1. We show that the random variable $Q = \mathrm{Z}_{\boldsymbol{h}}'\boldsymbol{V}^{-1}\mathrm{Z}_{\boldsymbol{h}}$ depends only on q and λ.

First, let $\boldsymbol{G}$ be a square root of $\boldsymbol{V}$ so that $\boldsymbol{V} = \boldsymbol{G}\boldsymbol{G}'$. Define $\mathrm{Z}_{\boldsymbol{h}}^* = \boldsymbol{G}^{-1}\mathrm{Z}_{\boldsymbol{h}} \sim \mathrm{N}(\boldsymbol{h}^*, \boldsymbol{I}_q)$ with $\boldsymbol{h}^* = \boldsymbol{G}^{-1}\boldsymbol{h}$. Note that $\lambda = \boldsymbol{h}'\boldsymbol{V}^{-1}\boldsymbol{h} = \boldsymbol{h}^{*\prime}\boldsymbol{h}^*$.

Second, construct an orthogonal $q \times q$ matrix $\boldsymbol{H} = [\boldsymbol{H}_1, \boldsymbol{H}_2]$ whose first column equals $\boldsymbol{H}_1 = \boldsymbol{h}^*\left(\boldsymbol{h}^{*\prime}\boldsymbol{h}^*\right)^{-1/2}$. Note that $\boldsymbol{H}_1'\boldsymbol{h}^* = \lambda^{1/2}$ and $\boldsymbol{H}_2'\boldsymbol{h}^* = \boldsymbol{0}$. Define $\mathrm{Z}_{\boldsymbol{h}}^{**} = \boldsymbol{H}'\mathrm{Z}_{\boldsymbol{h}}^* \sim \mathrm{N}(\boldsymbol{h}^{**}, \boldsymbol{I}_q)$ where

$$\boldsymbol{h}^{**} = \boldsymbol{H}'\boldsymbol{\delta}^* = \begin{pmatrix}\boldsymbol{H}_1'\boldsymbol{h}^* \\ \boldsymbol{H}_2'\boldsymbol{h}^*\end{pmatrix} = \begin{pmatrix}\lambda^{1/2} \\ \boldsymbol{0}\end{pmatrix}\begin{matrix}1 \\ q-1\end{matrix}.$$

Note that the distribution of $\mathrm{Z}_{\boldsymbol{h}}^{**}$ is only a function of λ and q.

Finally, observe that

$$\begin{aligned}Q &= \mathrm{Z}_{\boldsymbol{h}}'\boldsymbol{V}^{-1}\mathrm{Z}_{\boldsymbol{h}} \\ &= \mathrm{Z}_{\boldsymbol{h}}^{*\prime}\mathrm{Z}_{\boldsymbol{h}}^* \\ &= \mathrm{Z}_{\boldsymbol{h}}^{**\prime}\mathrm{Z}_{\boldsymbol{h}}^{**}\end{aligned}$$

which is only a function of $\mathrm{Z}_{\boldsymbol{h}}^{**}$ and thus its distribution only depends on λ and q. ■

Exercises

Exercise 8.1 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 $t_{k+1} = \hat{\beta}_{k+1}/s(\hat{\beta}_{k+1})$ is the t-ratio for $\hat{\beta}_{k+1}$ and

$$s(\hat{\beta}_{k+1}) = \left(s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{k+1,k+1}\right)^{1/2}$$

is the homoskedasticity-formula standard error.

Exercise 8.2 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 8.3 The data set `invest.dat` contains data on 565 U.S. firms extracted from Compustat for the year 1987. The variables, in order, are

• I_i	Investment to Capital Ratio (multiplied by 100).
• Q_i	Total Market Value to Asset Ratio (Tobin's Q).
• C_i	Cash Flow to Asset Ratio.
• D_i	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 8.4 In a paper in 1963, Marc Nerlove analyzed a cost function for 145 American electric companies. (The problem is discussed in Example 8.3 of Greene, section 1.7 of Hayashi, and the empirical exercise in Chapter 1 of Hayashi). The data file `nerlov.dat` contains his data. The variables are described on page 77 of Hayashi. Nerlov was interested in estimating a *cost function*: $TC = f(Q, PL, PF, PK)$.

(a) First estimate an unrestricted Cobb-Douglass specification

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + e_i. \tag{8.29}$$

Report parameter estimates and standard errors. You should obtain the same OLS estimates as in Hayashi's equation (1.7.7), but your standard errors may differ.

(b) What is the economic meaning of the restriction $\mathbb{H}_0 : \beta_3 + \beta_4 + \beta_5 = 1$?

(c) Estimate (8.29) by constrained least-squares imposing $\beta_3+\beta_4+\beta_5 = 1$. Report your parameter estimates and standard errors.

(d) Estimate (8.29) 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

Chapter 9

Regression Extensions

9.1 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 **non-linear regression**.

Example 9.1.1 *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 9.1.2 *Logistic Link Regression*

$$m(\boldsymbol{x},\boldsymbol{\theta}) = \Lambda\left(\boldsymbol{x}'\boldsymbol{\theta}\right)$$

where

$$\Lambda(u) = (1+\exp(-u))^{-1} \tag{9.1}$$

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 9.1.3 *Exponentially Transformed Regressors*

$$m(x,\boldsymbol{\theta}) = \theta_1 + \theta_2 \exp(\theta_3 x)$$

Example 9.1.4 *Power Transformation*

$$m(x,\boldsymbol{\theta}) = \theta_1 + \theta_2 x^{\theta_3}$$

with $x > 0$.

Example 9.1.5 *Box-Cox Transformed Regressors*

$$m(x,\boldsymbol{\theta}) = \theta_1 + \theta_2 x^{(\theta_3)}$$

where

$$x^{(\lambda)} = \left\{ \begin{array}{cc} \dfrac{x^\lambda - 1}{\lambda}, & \textit{if } \lambda > 0 \\ \log(x), & \textit{if } \lambda = 0 \end{array} \right\} \tag{9.2}$$

and $x > 0$. *The function (9.2) 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 9.1.6 *Continuous Threshold Regression*

$$m\left(x,\boldsymbol{\theta}\right)=\theta_{1}+\theta_{2}x+\theta_{3}\left(x-\theta_{4}\right)1\left(x>\theta_{4}\right)$$

Example 9.1.7 *Threshold Regression*

$$m\left(\boldsymbol{x},\boldsymbol{\theta}\right)=\left(\theta_{1}'\boldsymbol{x}_{1}\right)1\left(x_{2}<\theta_{3}\right)+\left(\theta_{2}'\boldsymbol{x}_{1}\right)1\left(x_{2}\geq\theta_{3}\right)$$

Example 9.1.8 *Smooth Transition*

$$m\left(\boldsymbol{x},\boldsymbol{\theta}\right)=\theta_{1}'\boldsymbol{x}_{1}+\left(\theta_{2}'\boldsymbol{x}_{1}\right)\Lambda\left(\frac{x_{2}-\theta_{3}}{\theta_{4}}\right)$$

where $\Lambda(u)$ *is the logit function (9.1).*

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 $m\left(\boldsymbol{x},\boldsymbol{\theta}\right)$ 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

$$S_{n}(\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^{n}\left(y_{i}-m\left(\boldsymbol{x}_{i},\boldsymbol{\theta}\right)\right)^{2}.$$

When the regression function is nonlinear, we call this the **nonlinear least squares** (NLLS) estimator. The NLLS residuals are $\hat{e}_{i}=y_{i}-m\left(\boldsymbol{x}_{i},\widehat{\boldsymbol{\theta}}\right)$.

One motivation for the choice of NLLS as the estimation method is that the parameter $\boldsymbol{\theta}$ is the solution to the population problem $\min_{\boldsymbol{\theta}}\mathbb{E}\left(y_{i}-m\left(\boldsymbol{x}_{i},\boldsymbol{\theta}\right)\right)^{2}$

Since sum-of-squared-errors function $S_{n}(\boldsymbol{\theta})$ 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

$$\mathbf{0}=\sum_{i=1}^{n}\boldsymbol{m}_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{i},\widehat{\boldsymbol{\theta}}\right)\hat{e}_{i}\tag{9.3}$$

where

$$\boldsymbol{m}_{\boldsymbol{\theta}}\left(\boldsymbol{x},\boldsymbol{\theta}\right)=\frac{\partial}{\partial\boldsymbol{\theta}}m\left(\boldsymbol{x},\boldsymbol{\theta}\right).$$

Theorem 9.1.1 ***Asymptotic Distribution of NLLS Estimator***
If the model is identified and $m\left(\boldsymbol{x},\boldsymbol{\theta}\right)$ *is differentiable with respect to* $\boldsymbol{\theta}$,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\theta}}\right)$$

$$\boldsymbol{V}_{\boldsymbol{\theta}}=\left(\mathbb{E}\left(\boldsymbol{m}_{\boldsymbol{\theta}i}\boldsymbol{m}_{\boldsymbol{\theta}i}'\right)\right)^{-1}\left(\mathbb{E}\left(\boldsymbol{m}_{\boldsymbol{\theta}i}\boldsymbol{m}_{\boldsymbol{\theta}i}'e_{i}^{2}\right)\right)\left(\mathbb{E}\left(\boldsymbol{m}_{\boldsymbol{\theta}i}\boldsymbol{m}_{\boldsymbol{\theta}i}'\right)\right)^{-1}$$

where $\boldsymbol{m}_{\boldsymbol{\theta}i}=\boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{x}_{i},\boldsymbol{\theta}_{0})$.

Based on Theorem 9.1.1, an estimate of the asymptotic variance $\boldsymbol{V_\theta}$ is

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}\hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}\hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}'\hat{e}_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}\hat{\boldsymbol{m}}_{\boldsymbol{\theta}i}'\right)^{-1}$$

where $\hat{\boldsymbol{m}}_{\boldsymbol{\theta}i} = \boldsymbol{m}_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})$ and $\hat{e}_i = y_i - m(\boldsymbol{x}_i, \widehat{\boldsymbol{\theta}})$.

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 9.1.1 (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 $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$. Proving that nonlinear estimators are consistent is more challenging than for linear estimators. We sketch the main argument. The idea is that $\hat{\boldsymbol{\theta}}$ minimizes the sample criterion function $S_n(\boldsymbol{\theta})$, which (for any $\boldsymbol{\theta}$) converges in probability to the mean-squared error function $\mathbb{E}\left(y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta})\right)^2$. Thus it seems reasonable that the minimizer $\hat{\boldsymbol{\theta}}$ will converge in probability to $\boldsymbol{\theta}_0$, the minimizer of $\mathbb{E}\left(y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta})\right)^2$. It turns out that to show this rigorously, we need to show that $S_n(\boldsymbol{\theta})$ converges *uniformly* to its expectation $\mathbb{E}\left(y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta})\right)^2$, which means that the maximum discrepancy must converge in probability to zero, to exclude the possibility that $S_n(\boldsymbol{\theta})$ 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 differentiabel in $\boldsymbol{\theta}$. For a complete treatment of the theory of optimization estimators see Newey and McFadden (1994).

Since $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$, $\hat{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}_0$ for n large, so the minimization of $S_n(\boldsymbol{\theta})$ 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

$$\begin{aligned} y_i - m(\boldsymbol{x}_i, \boldsymbol{\theta}) &\simeq (e_i + m(\boldsymbol{x}_i, \boldsymbol{\theta}_0)) - \big(m(\boldsymbol{x}_i, \boldsymbol{\theta}_0) + \boldsymbol{m}_{\boldsymbol{\theta}i}'(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\big) \\ &= e_i - \boldsymbol{m}_{\boldsymbol{\theta}i}'(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ &= y_i^0 - \boldsymbol{m}_{\boldsymbol{\theta}i}'\boldsymbol{\theta}. \end{aligned}$$

Hence the sum of squared errors function is

$$S_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y_i - m\left(\boldsymbol{x}_i, \boldsymbol{\theta}\right)\right)^2 \simeq \sum_{i=1}^{n} \left(y_i^0 - \boldsymbol{m}_{\boldsymbol{\theta} i}' \boldsymbol{\theta}\right)^2$$

and the right-hand-side is the SSE 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.

9.2 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 (4.13) introduced in Section 4.6.1. The GLS estimator is sometimes called the Aitken estimator. The GLS estimator (9.2) is infeasible since the matrix $\boldsymbol{D}$ is unknown. A feasible GLS (FGLS) estimator replaces the unknown $\boldsymbol{D}$ with an estimate $\hat{\boldsymbol{D}} = \operatorname{diag}\{\hat{\sigma}_1^2, ..., \hat{\sigma}_n^2\}$. 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

$$\begin{aligned} \eta_i &= \alpha_0 + \boldsymbol{z}_{1i}' \boldsymbol{\alpha}_1 + \xi_i \\ \mathbb{E}\left(\xi_i \mid \boldsymbol{x}_i\right) &= 0. \end{aligned} \tag{9.4}$$

This regression error ξ_i is generally heteroskedastic and has the conditional variance

$$\begin{aligned} \operatorname{var}\left(\xi_i \mid \boldsymbol{x}_i\right) &= \operatorname{var}\left(e_i^2 \mid \boldsymbol{x}_i\right) \\ &= \mathbb{E}\left(\left(e_i^2 - \mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right)\right)^2 \mid \boldsymbol{x}_i\right) \\ &= \mathbb{E}\left(e_i^4 \mid \boldsymbol{x}_i\right) - \left(\mathbb{E}\left(e_i^2 \mid \boldsymbol{x}_i\right)\right)^2. \end{aligned}$$

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

$$\sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}_{\boldsymbol{\alpha}}\right)$$

where

$$\boldsymbol{V}_{\boldsymbol{\alpha}} = \left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\xi_i^2\right)\left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}. \tag{9.5}$$

While e_i is not observed, we have the OLS residual $\hat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} = e_i - \boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Thus

$$\begin{aligned}\phi_i &\equiv \hat{\eta}_i - \eta_i \\ &= \hat{e}_i^2 - e_i^2 \\ &= -2e_i\boldsymbol{x}_i'\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{x}_i\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).\end{aligned}$$

And then

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{z}_i\phi_i &= \frac{-2}{n}\sum_{i=1}^n \boldsymbol{z}_i e_i\boldsymbol{x}_i'\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) + \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{x}_i\boldsymbol{x}_i'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\sqrt{n} \\ &\xrightarrow{p} \boldsymbol{0}\end{aligned}$$

Let

$$\widetilde{\boldsymbol{\alpha}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\hat{\boldsymbol{\eta}} \tag{9.6}$$

be from OLS regression of $\hat{\eta}_i$ on $\boldsymbol{z}_i$. Then

$$\begin{aligned}\sqrt{n}\left(\widetilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) &= \sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right) + \left(n^{-1}\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}n^{-1/2}\boldsymbol{Z}'\boldsymbol{\phi} \\ &\xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\alpha}}\right)\end{aligned} \tag{9.7}$$

Thus the fact that η_i is replaced with $\hat{\eta}_i$ is asymptotically irrelevant. We call (9.6) 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

$$\tilde{\sigma}_i^2 = \widetilde{\boldsymbol{\alpha}}'\boldsymbol{z}_i. \tag{9.8}$$

Suppose that $\tilde{\sigma}_i^2 > 0$ for all i. Then set

$$\widetilde{\boldsymbol{D}} = \mathrm{diag}\{\tilde{\sigma}_1^2, ..., \tilde{\sigma}_n^2\}$$

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 (9.4), there is no guarantee that $\tilde{\sigma}_i^2 > 0$ for all i. If $\tilde{\sigma}_i^2 < 0$ for some i, then the FGLS estimator is not well defined. Furthermore, if $\tilde{\sigma}_i^2 \approx 0$ 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

$$\overline{\sigma}_i^2 = \max[\tilde{\sigma}_i^2, c\hat{\sigma}^2]$$

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 9.2.1 *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 9.2.1, the natural estimator of the asymptotic variance of $\widetilde{\boldsymbol{\beta}}$ is

$$\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}^0=\left(\frac{1}{n}\sum_{i=1}^n\tilde{\sigma}_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}=\left(\frac{1}{n}\boldsymbol{X}'\tilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}.$$

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 (9.4) 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 $\boldsymbol{z}_i$. $\tilde{\boldsymbol{\beta}}$ should perhaps be called a quasi-FGLS estimator of $\boldsymbol{\beta}$. Its asymptotic variance is not that given in Theorem 9.2.1. 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

$$\begin{aligned}\widetilde{\boldsymbol{V}}_{\boldsymbol{\beta}}&=\left(\frac{1}{n}\sum_{i=1}^n\tilde{\sigma}_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n\tilde{\sigma}_i^{-4}\hat{e}_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\frac{1}{n}\sum_{i=1}^n\tilde{\sigma}_i^{-2}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}\\&=\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{D}}^{-1}\widehat{\boldsymbol{D}}\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\widetilde{\boldsymbol{D}}^{-1}\boldsymbol{X}\right)^{-1}\end{aligned}$$

where $\widehat{\boldsymbol{D}}=\operatorname{diag}\{\hat{e}_1^2,...,\hat{e}_n^2\}$. 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.

9.3 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 (9.4). We may therefore test this hypothesis by the estimation (9.6) 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 (9.5) for $\tilde{\boldsymbol{\alpha}}$ simplifies to

$$\boldsymbol{V}_{\boldsymbol{\alpha}} = \left(\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\right)^{-1}\mathbb{E}\left(\xi_i^2\right). \tag{9.9}$$

Hence the standard test of $\mathbb{H}_0$ is a classic F (or Wald) test for exclusion of all regressors from the skedastic regression (9.6). The asymptotic distribution (9.7) and the asymptotic variance (9.9) under independence show that this test has an asymptotic chi-square distribution.

Theorem 9.3.1 *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

9.4 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 $y_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + \hat{e}_i$ 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 $y_i = \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}} + \boldsymbol{z}_i'\tilde{\boldsymbol{\gamma}} + \tilde{e}_i$ 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 $\hat{y}_i = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$. Now let

$$\boldsymbol{z}_i = \begin{pmatrix} \hat{y}_i^2 \\ \vdots \\ \hat{y}_i^m \end{pmatrix}$$

be a $(m-1)$-vector of powers of $\hat{y}_i$. Then run the auxiliary regression

$$y_i = \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}} + \boldsymbol{z}_i'\widetilde{\boldsymbol{\gamma}} + \tilde{e}_i \tag{9.10}$$

by OLS, and form the Wald statistic W_n for $\boldsymbol{\gamma} = \mathbf{0}$. It is easy (although somewhat tedious) to show that under the null hypothesis, $W_n \xrightarrow{d} \chi^2_{m-1}$. Thus the null is rejected at the $\alpha\%$ level if W_n exceeds the upper $\alpha\%$ tail 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 (9.10) may be written as

$$y_i = \boldsymbol{x}_i'\widetilde{\boldsymbol{\beta}} + \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right)^2 \tilde{\gamma}_1 + \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right)^3 \tilde{\gamma}_2 + \cdots \left(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right)^m \tilde{\gamma}_{m-1} + \tilde{e}_i$$

which has essentially approximated $G(\cdot)$ by a m'th order polynomial

9.5 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

$$\theta = \underset{\theta}{\text{argmin}}\, \mathbb{E}\,|y - \theta| \tag{9.11}$$

and

$$\mathbb{E}\,\text{sgn}\,(y - \theta) = 0$$

where

$$\text{sgn}\,(u) = \begin{cases} 1 & \text{if } u \ge 0 \\ -1 & \text{if } u < 0 \end{cases}$$

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 $\frac{1}{n}\sum_{i=1}^n |y_i - \theta|$, and the third definition is the solution to the moment equation $\frac{1}{n}\sum_{i=1}^n \text{sgn}\,(y_i - \theta)$. 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 $m(\boldsymbol{x}) = \text{med}\,(y \mid \boldsymbol{x})$ denote the conditional median of y given $\boldsymbol{x}$. The linear median regression model takes the form

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$$
$$\text{med}\,(e_i \mid \boldsymbol{x}_i) = 0$$

In this model, the linear function $\text{med}\,(y_i \mid \boldsymbol{x}_i = \boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$ 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

- $\Pr(y_i \leq \boldsymbol{x}'\boldsymbol{\beta} \mid \boldsymbol{x}_i = \boldsymbol{x}) = \Pr(y_i > \boldsymbol{x}'\boldsymbol{\beta} \mid \boldsymbol{x}_i = \boldsymbol{x}) = .5$
- $\mathbb{E}\left(\operatorname{sgn}\left(e_i\right) \mid \boldsymbol{x}_i\right) = 0$
- $\mathbb{E}\left(\boldsymbol{x}_i \operatorname{sgn}\left(e_i\right)\right) = 0$
- $\boldsymbol{\beta} = \min_{\boldsymbol{\beta}} \mathbb{E}\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$

These facts motivate the following estimator. Let

$$LAD_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$$

be the average of absolute deviations. The **least absolute deviations** (LAD) estimator of $\boldsymbol{\beta}$ minimizes this function

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, LAD_n(\boldsymbol{\beta})$$

Equivalently, it is a solution to the moment condition

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \operatorname{sgn}\left(y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}\right) = 0. \tag{9.12}$$

The LAD estimator has an asymptotic normal distribution.

Theorem 9.5.1 ***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

$$V = \frac{1}{4}\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}\left(\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i' f\left(0 \mid \boldsymbol{x}_i\right)\right)\right)^{-1}$$

and $f\left(e \mid \boldsymbol{x}\right)$ *is the conditional density of* e_i *given* $\boldsymbol{x}_i = \boldsymbol{x}$.

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

$$\boldsymbol{V} = \frac{\left(\mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\right)^{-1}}{4f\left(0\right)^2} \tag{9.13}$$

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 (9.13). 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 20. While a complete proof of Theorem 9.5.1 is advanced, we provide a sketch here for completeness.

Proof of Theorem 9.5.1: 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 $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$. 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, $LAD_n(\boldsymbol{\beta}) \xrightarrow{p} \mathbb{E}\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$. 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 $LAD_n(\boldsymbol{\beta})$, converges in probability to $\boldsymbol{\beta}_0$, the minimizer of $\mathbb{E}\left|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right|$.

Since $\operatorname{sgn}(a) = 1 - 2 \cdot 1\left(a \leq 0\right)$, (9.12) 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 $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta})$. We need three preliminary results. First, since $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0) = 0$ and $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0)\boldsymbol{g}_i(\boldsymbol{\beta}_0)' = \mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'$, we can apply the central limit theorem (Theorem 5.7.1) and find that

$$\sqrt{n}\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) = n^{-1/2}\sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\right).$$

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

$$\begin{aligned}
\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) &\simeq \left(\frac{\partial}{\partial\boldsymbol{\beta}'}\boldsymbol{g}(\boldsymbol{\beta}_0)\right)^{-1}\sqrt{n}\boldsymbol{g}(\hat{\boldsymbol{\beta}}) \\
&= \left(-2\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'f\left(0 \mid \boldsymbol{x}_i\right)\right]\right)^{-1}\sqrt{n}\left(\boldsymbol{g}(\hat{\boldsymbol{\beta}}) - \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})\right) \\
&\simeq \frac{1}{2}\left(\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'f\left(0 \mid \boldsymbol{x}_i\right)\right]\right)^{-1}\sqrt{n}\left(\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) - \boldsymbol{g}(\boldsymbol{\beta}_0)\right) \\
&\xrightarrow{d} \frac{1}{2}\left(\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i'f\left(0 \mid \boldsymbol{x}_i\right)\right]\right)^{-1}\mathrm{N}\left(\mathbf{0}, \mathbb{E}\boldsymbol{x}_i\boldsymbol{x}_i'\right) \\
&= \mathrm{N}\left(\mathbf{0}, \boldsymbol{V}\right).
\end{aligned}$$

The third line follows from an asymptotic empirical process argument and the fact that $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.

9.6 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 $F(Q_\tau) = \tau$, 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

$$Q_\tau = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}\rho_\tau (U - \theta). \tag{9.14}$$

where $\rho_\tau(q)$ is the piecewise linear function

$$\begin{aligned}\rho_\tau(q) &= \begin{cases} -q(1-\tau) & q < 0 \\ q\tau & q \geq 0 \end{cases} \\ &= q(\tau - 1(q<0)). \end{aligned} \tag{9.15}$$

This generalizes representation (9.11) for the median to all quantiles.

For the random variables $(y_i, \boldsymbol{x}_i)$ with conditional distribution function $F(y \mid \boldsymbol{x})$ the conditional quantile function $q_\tau(\boldsymbol{x})$ is

$$Q_\tau(\boldsymbol{x}) = \inf\{y : F(y \mid \boldsymbol{x}) \geq \tau\}.$$

Again, when $F(y \mid \boldsymbol{x})$ is continuous and strictly monotonic in y, then $F(Q_\tau(\boldsymbol{x}) \mid \boldsymbol{x}) = \tau$. 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 (9.14), the quantile regression estimator $\widehat{\boldsymbol{\beta}}_\tau$ for $\boldsymbol{\beta}_\tau$ solves the minimization problem

$$\widehat{\boldsymbol{\beta}}_\tau = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, S_n^\tau(\boldsymbol{\beta})$$

where

$$S_n^\tau(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \rho_\tau\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)$$

and $\rho_\tau(q)$ is defined in (9.15).

Since the quantile regression criterion function $S_n^\tau(\boldsymbol{\beta})$ 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 9.5.1.

Theorem 9.6.1 ***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

$$\boldsymbol{V}_{\tau}=\tau(1-\tau)\left(\mathbb{E}\left(\boldsymbol{x}_{i} \boldsymbol{x}_{i}^{\prime} f\left(0 \mid \boldsymbol{x}_{i}\right)\right)\right)^{-1}\left(\mathbb{E} \boldsymbol{x}_{i} \boldsymbol{x}_{i}^{\prime}\right)\left(\mathbb{E}\left(\boldsymbol{x}_{i} \boldsymbol{x}_{i}^{\prime} f\left(0 \mid \boldsymbol{x}_{i}\right)\right)\right)^{-1}$$

and $f(e \mid \boldsymbol{x})$ is the conditional density of e_i given $\boldsymbol{x}_i=\boldsymbol{x}$.

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 $f(0 \mid \boldsymbol{x}_i)=f(0)$, the unconditional density of e_i at 0, and we have the simplification

$$\boldsymbol{V}_{\tau}=\frac{\tau(1-\tau)}{f(0)^{2}}\left(\mathbb{E}\left(\boldsymbol{x}_{i} \boldsymbol{x}_{i}^{\prime}\right)\right)^{-1}.$$

A recent monograph on the details of quantile regression is Koenker (2005).

Exercises

Exercise 9.1 Suppose that $y_i = g(\boldsymbol{x}_i, \boldsymbol{\theta}) + e_i$ with $\mathbb{E}(e_i \mid \boldsymbol{x}_i) = 0$, $\hat{\boldsymbol{\theta}}$ is the NLLS estimator, and $\hat{\boldsymbol{V}}$ is the estimate of $\operatorname{var}\left(\hat{\boldsymbol{\theta}}\right)$. 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 9.2 In Exercise 8.4, you estimated a cost function on a cross-section of electric companies. The equation you estimated was

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log PL_i + \beta_4 \log PK_i + \beta_5 \log PF_i + e_i. \tag{9.16}$$

(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 (9.16) 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 9.3 The data file `cps78.dat` contains 550 observations on 20 variables taken from the May 1978 current population survey. Variables are listed in the file `cps78.pdf`. The goal of the exercise is to estimate a model for the log of earnings (variable LNWAGE) as a function of the conditioning variables.

(a) Start by an OLS regression of LNWAGE on the other variables. Report coefficient estimates and standard errors.

(b) Consider augmenting the model by squares and/or cross-products of the conditioning variables. Estimate your selected model and report the results.

(c) Are there any variables which seem to be unimportant as a determinant of wages? You may re-estimate the model without these variables, if desired.

(d) Test whether the error variance is different for men and women. Interpret.

(e) Test whether the error variance is different for whites and nonwhites. Interpret.

(f) Construct a model for the conditional variance. Estimate such a model, test for general heteroskedasticity and report the results.

(g) Using this model for the conditional variance, re-estimate the model from part (c) using FGLS. Report the results.

(h) Do the OLS and FGLS estimates differ greatly? Note any interesting differences.

(i) Compare the estimated standard errors. Note any interesting differences.

Exercise 9.4 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 $m\left(\boldsymbol{x}\right) = \operatorname{med}(y_i \mid \boldsymbol{x}_i)$.

Exercise 9.5 Define

$$g(u) = \tau - 1\left(u < 0\right)$$

where $1\left(\cdot\right)$ 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 9.6 Verify equation (9.14).

Chapter 10

The Bootstrap

10.1 Definition of the Bootstrap

Let F denote a distribution function for the population of observations $(y_i, \boldsymbol{x}_i)$. Let

$$T_n = T_n\left((y_1, \boldsymbol{x}_1), ..., (y_n, \boldsymbol{x}_n), F\right)$$

be a statistic of interest, for example an estimator $\hat{\theta}$ or a t-statistic $\left(\hat{\theta} - \theta\right)/s(\hat{\theta})$. Note that we write T_n as possibly a function of F. For example, the t-statistic is a function of the parameter θ which itself is a function of F.

The exact CDF of T_n when the data are sampled from the distribution F is

$$G_n(u, F) = \Pr(T_n \leq u \mid F)$$

In general, $G_n(u, F)$ depends on F, meaning that G changes as F 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_n 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 F_n (one choice is discussed in the next section). Plugged into $G_n(u, F)$ we obtain

$$G_n^*(u) = G_n(u, F_n). \tag{10.1}$$

We call G_n^* the bootstrap distribution. Bootstrap inference is based on $G_n^*(u)$.

Let $(y_i^*, \boldsymbol{x}_i^*)$ denote random variables with the distribution F_n. A random sample from this distribution is called the bootstrap data. The statistic $T_n^* = T_n\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), F_n\right)$ constructed on this sample is a random variable with distribution G_n^*. That is, $\Pr(T_n^* \leq u) = G_n^*(u)$. We call T_n^* the bootstrap statistic. The distribution of T_n^* is identical to that of T_n when the true CDF is F_n rather than F.

The bootstrap distribution is itself random, as it depends on the sample through the estimator F_n.

In the next sections we describe computation of the bootstrap distribution.

10.2 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

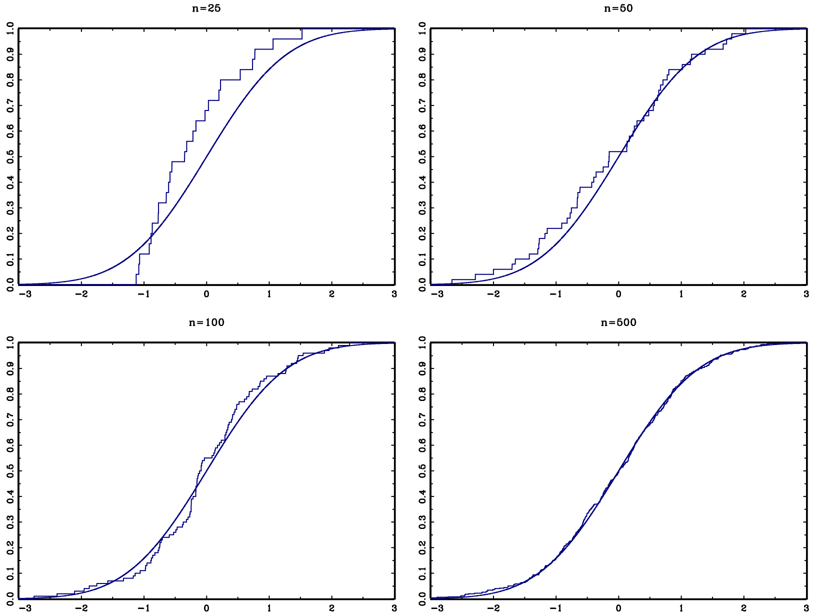


Figure 10.1: Empirical Distribution Functions

sample moment:

$$F_n(y, \boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} 1(y_i \le y)\, 1(\boldsymbol{x}_i \le \boldsymbol{x}). \tag{10.2}$$

$F_n(y, \boldsymbol{x})$ is called the empirical distribution function (EDF). F_n is a nonparametric estimate of F. Note that while F may be either discrete or continuous, F_n is by construction a step function.

The EDF is a consistent estimator of the CDF. To see this, note that for any $(y, \boldsymbol{x})$, $1(y_i \le y)\, 1(\boldsymbol{x}_i \le \boldsymbol{x})$ is an iid random variable with expectation $F(y, \boldsymbol{x})$. Thus by the WLLN (Theorem 5.4.2), $F_n(y, \boldsymbol{x}) \xrightarrow{p} F(y, \boldsymbol{x})$. Furthermore, by the CLT (Theorem 5.7.1),

$$\sqrt{n}\,(F_n(y, \boldsymbol{x}) - F(y, \boldsymbol{x})) \xrightarrow{d} \mathrm{N}(0, F(y, \boldsymbol{x})(1 - F(y, \boldsymbol{x}))).$$

To see the effect of sample size on the EDF, in the Figure below, 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 F_n. That is,

$$\Pr(y_i^* \le y, \boldsymbol{x}_i^* \le \boldsymbol{x}) = F_n(y, \boldsymbol{x}).$$

We can easily calculate the moments of functions of $(y_i^*, \boldsymbol{x}_i^*)$:

$$\begin{aligned}\mathbb{E}h(y_i^*, \boldsymbol{x}_i^*) &= \int h(y, \boldsymbol{x})dF_n(y, \boldsymbol{x}) \\ &= \sum_{i=1}^{n} h(y_i, \boldsymbol{x}_i)\Pr(y_i^* = y_i, \boldsymbol{x}_i^* = \boldsymbol{x}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} h(y_i, \boldsymbol{x}_i),\end{aligned}$$

the empirical sample average.

10.3 Nonparametric Bootstrap

The **nonparametric bootstrap** is obtained when the bootstrap distribution (10.1) is defined using the EDF (10.2) as the estimate F_n of F.

Since the EDF F_n 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 $T_n^* = T_n\left((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*), F_n\right)$ 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_n is a function of F, it is typically through dependence on a parameter. For example, the t-ratio $\left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ depends on θ. As the bootstrap statistic replaces F with F_n, it similarly replaces θ with θ_n, the value of θ implied by F_n. Typically $\theta_n = \hat{\theta}$, the parameter estimate. (When in doubt use $\hat{\theta}$.)

Sampling from the EDF is particularly easy. Since F_n 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.

10.4 Bootstrap Estimation of Bias and Variance

The bias of $\hat{\theta}$ is $\tau_n = \mathbb{E}(\hat{\theta} - \theta_0)$. Let $T_n(\theta) = \hat{\theta} - \theta$. Then $\tau_n = \mathbb{E}(T_n(\theta_0))$. The bootstrap counterparts are $\hat{\theta}^* = \hat{\theta}((y_1^*, \boldsymbol{x}_1^*), ..., (y_n^*, \boldsymbol{x}_n^*))$ and $T_n^* = \hat{\theta}^* - \theta_n = \hat{\theta}^* - \hat{\theta}$. The bootstrap estimate of τ_n is

$$\tau_n^* = \mathbb{E}(T_n^*).$$

If this is calculated by the simulation described in the previous section, the estimate of τ_n^* is

$$\begin{aligned}\hat{\tau}_n^* &= \frac{1}{B}\sum_{b=1}^{B} T_{nb}^* \\ &= \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_b^* - \hat{\theta} \\ &= \overline{\hat{\theta}^*} - \hat{\theta}.\end{aligned}$$

If $\hat{\theta}$ is biased, it might be desirable to construct a biased-corrected estimator (one with reduced bias). Ideally, this would be

$$\tilde{\theta} = \hat{\theta} - \tau_n,$$

but τ_n is unknown. The (estimated) bootstrap biased-corrected estimator is

$$\begin{aligned}\tilde{\theta}^* &= \hat{\theta} - \hat{\tau}_n^* \\ &= \hat{\theta} - \left(\overline{\hat{\theta}^*} - \hat{\theta}\right) \\ &= 2\hat{\theta} - \overline{\hat{\theta}^*}.\end{aligned}$$

Note, in particular, that the biased-corrected estimator is *not* $\overline{\hat{\theta}^*}$. Intuitively, the bootstrap makes the following experiment. Suppose that $\hat{\theta}$ is the truth. Then what is the average value of $\hat{\theta}$ calculated from such samples? The answer is $\overline{\hat{\theta}^*}$. If this is lower than $\hat{\theta}$, this suggests that the estimator is *downward-biased*, so a biased-corrected estimator of θ should be *larger* than $\hat{\theta}$, and the best guess is the difference between $\hat{\theta}$ and $\overline{\hat{\theta}^*}$. Similarly if $\overline{\hat{\theta}^*}$ is higher than $\hat{\theta}$, then the estimator is upward-biased and the biased-corrected estimator should be lower than $\hat{\theta}$.

Let $T_n = \hat{\theta}$. The variance of $\hat{\theta}$ is

$$V_n = \mathbb{E}(T_n - \mathbb{E}T_n)^2.$$

Let $T_n^* = \hat{\theta}^*$. It has variance

$$V_n^* = \mathbb{E}(T_n^* - \mathbb{E}T_n^*)^2.$$

The simulation estimate is

$$\hat{V}_n^* = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b^* - \overline{\hat{\theta}^*}\right)^2.$$

A bootstrap standard error for $\hat{\theta}$ is the square root of the bootstrap estimate of variance, $s^*(\hat{\theta}) = \sqrt{\hat{V}_n^*}$.

While this standard error may be calculated and reported, it is not clear if it is useful. The primary use of asymptotic standard errors is to construct asymptotic confidence intervals, which are based on the asymptotic normal approximation to the t-ratio. However, the use of the bootstrap presumes that such asymptotic approximations might be poor, in which case the normal approximation is suspected. It appears superior to calculate bootstrap confidence intervals, and we turn to this next.

10.5 Percentile Intervals

For a distribution function $G_n(u, F)$, let $q_n(\alpha, F)$ denote its quantile function. This is the function which solves

$$G_n(q_n(\alpha, F), F) = \alpha.$$

[When $G_n(u, F)$ is discrete, $q_n(\alpha, F)$ may be non-unique, but we will ignore such complications.] Let $q_n(\alpha)$ denote the quantile function of the true sampling distribution, and $q_n^*(\alpha) = q_n(\alpha, F_n)$ denote the quantile function of the bootstrap distribution. Note that this function will change depending on the underlying statistic T_n whose distribution is G_n.

Let $T_n = \hat{\theta}$, an estimate of a parameter of interest. In $(1-\alpha)\%$ of samples, $\hat{\theta}$ lies in the region $[q_n(\alpha/2), q_n(1-\alpha/2)]$. This motivates a confidence interval proposed by Efron:

$$C_1 = [q_n^*(\alpha/2), \quad q_n^*(1-\alpha/2)].$$

This is often called the *percentile confidence interval.*

Computationally, the quantile $q_n^*(\alpha)$ is estimated by $\hat{q}_n^*(\alpha)$, the α'th sample quantile of the simulated statistics $\{T_{n1}^*, ..., T_{nB}^*\}$, as discussed in the section on Monte Carlo simulation. The $(1-\alpha)\%$ Efron percentile interval is then $[\hat{q}_n^*(\alpha/2), \quad \hat{q}_n^*(1-\alpha/2)]$.

The interval 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 $[f(q_n^*(\alpha/2)), \quad f(q_n^*(1-\alpha/2))]$, which is a naturally good property.

However, as we show now, C_1 is in a deep sense very poorly motivated.

It will be useful if we introduce an alternative definition of C_1. Let $T_n(\theta) = \hat{\theta} - \theta$ and let $q_n(\alpha)$ be the quantile function of its distribution. (These are the original quantiles, with θ subtracted.) Then C_1 can alternatively be written as

$$C_1 = [\hat{\theta} + q_n^*(\alpha/2), \quad \hat{\theta} + q_n^*(1-\alpha/2)].$$

This is a bootstrap estimate of the "ideal" confidence interval

$$C_1^0 = [\hat{\theta} + q_n(\alpha/2), \quad \hat{\theta} + q_n(1-\alpha/2)].$$

The latter has coverage probability

$$\begin{aligned}
\Pr\left(\theta_0 \in C_1^0\right) &= \Pr\left(\hat{\theta} + q_n(\alpha/2) \le \theta_0 \le \hat{\theta} + q_n(1-\alpha/2)\right) \\
&= \Pr\left(-q_n(1-\alpha/2) \le \hat{\theta} - \theta_0 \le -q_n(\alpha/2)\right) \\
&= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1-\alpha/2), F_0)
\end{aligned}$$

which generally is not $1-\alpha$! There is one important exception. If $\hat{\theta} - \theta_0$ has a symmetric distribution about 0, then $G_n(-u, F_0) = 1 - G_n(u, F_0)$, so

$$\begin{aligned}
\Pr\left(\theta_0 \in C_1^0\right) &= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1-\alpha/2), F_0) \\
&= (1 - G_n(q_n(\alpha/2), F_0)) - (1 - G_n(q_n(1-\alpha/2), F_0)) \\
&= \left(1 - \frac{\alpha}{2}\right) - \left(1 - \left(1 - \frac{\alpha}{2}\right)\right) \\
&= 1 - \alpha
\end{aligned}$$

and this idealized confidence interval is accurate. Therefore, C_1^0 and C_1 are designed for the case that $\hat{\theta}$ has a symmetric distribution about θ_0.

When $\hat{\theta}$ does not have a symmetric distribution, 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 $f(\hat{\theta})$ is symmetrically distributed about $f(\theta_0)$, 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.

Let $T_n(\theta) = \hat{\theta} - \theta$. Then

$$\begin{aligned}
1 - \alpha &= \Pr\left(q_n(\alpha/2) \le T_n(\theta_0) \le q_n(1-\alpha/2)\right) \\
&= \Pr\left(\hat{\theta} - q_n(1-\alpha/2) \le \theta_0 \le \hat{\theta} - q_n(\alpha/2)\right),
\end{aligned}$$

so an exact $(1-\alpha)\%$ confidence interval for θ_0 would be

$$C_2^0 = [\hat{\theta} - q_n(1-\alpha/2), \quad \hat{\theta} - q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_2 = [\hat{\theta} - q_n^*(1-\alpha/2), \quad \hat{\theta} - q_n^*(\alpha/2)].$$

Notice that generally this is very different from the Efron interval C_1! They coincide in the special case that $G_n^*(u)$ is symmetric about $\hat{\theta}$, but otherwise they differ.

Computationally, this interval can be estimated from a bootstrap simulation by sorting the bootstrap statistics $T_n^* = \left(\hat{\theta}^* - \hat{\theta}\right)$, which are centered at the sample estimate $\hat{\theta}$. These are sorted to yield the quantile estimates $\hat{q}_n^*(.025)$ and $\hat{q}_n^*(.975)$. The 95% confidence interval is then $[\hat{\theta} - \hat{q}_n^*(.975), \quad \hat{\theta} - \hat{q}_n^*(.025)]$.

This confidence interval is discussed in most theoretical treatments of the bootstrap, but is not widely used in practice.

10.6 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 $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ and reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $T_n(\theta_0) < c$, where c would be selected so that

$$\Pr\left(T_n(\theta_0) < c\right) = \alpha.$$

Thus $c = q_n(\alpha)$. Since this is unknown, a bootstrap test replaces $q_n(\alpha)$ with the bootstrap estimate $q_n^*(\alpha)$, and the test rejects if $T_n(\theta_0) < q_n^*(\alpha)$.

Similarly, if the alternative is $\mathbb{H}_1 : \theta > \theta_0$, the bootstrap test rejects if $T_n(\theta_0) > q_n^*(1-\alpha)$.

Computationally, these critical values can be estimated from a bootstrap simulation by sorting the bootstrap t-statistics $T_n^* = \left(\hat{\theta}^* - \hat{\theta}\right)/s(\hat{\theta}^*)$. Note, and this is important, that the bootstrap test statistic is centered at the estimate $\hat{\theta}$, and the standard error $s(\hat{\theta}^*)$ is calculated on the bootstrap sample. These t-statistics are sorted to find the estimated quantiles $\hat{q}_n^*(\alpha)$ and/or $\hat{q}_n^*(1-\alpha)$.

Let $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$. Then taking the intersection of two one-sided intervals,

$$\begin{aligned}
1 - \alpha &= \Pr\left(q_n(\alpha/2) \leq T_n(\theta_0) \leq q_n(1-\alpha/2)\right) \\
&= \Pr\left(q_n(\alpha/2) \leq \left(\hat{\theta} - \theta_0\right)/s(\hat{\theta}) \leq q_n(1-\alpha/2)\right) \\
&= \Pr\left(\hat{\theta} - s(\hat{\theta})q_n(1-\alpha/2) \leq \theta_0 \leq \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)\right),
\end{aligned}$$

so an exact $(1-\alpha)\%$ confidence interval for θ_0 would be

$$C_3^0 = [\hat{\theta} - s(\hat{\theta})q_n(1-\alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n(\alpha/2)].$$

This motivates a bootstrap analog

$$C_3 = [\hat{\theta} - s(\hat{\theta})q_n^*(1-\alpha/2), \quad \hat{\theta} - s(\hat{\theta})q_n^*(\alpha/2)].$$

This is often called a *percentile-t confidence interval.* It is *equal-tailed* or *central* since the probability that θ_0 is below the left endpoint approximately equals the probability that θ_0 is above the right endpoint, each $\alpha/2$.

Computationally, this is based on the critical values from the one-sided hypothesis tests, discussed above.

10.7 Symmetric Percentile-t Intervals

Suppose we want to test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ at size α. We would set $T_n(\theta) = \left(\hat{\theta} - \theta\right)/s(\hat{\theta})$ and reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $|T_n(\theta_0)| > c$, where c would be selected so that

$$\Pr\left(|T_n(\theta_0)| > c\right) = \alpha.$$

Note that

$$\begin{aligned}\Pr\left(|T_n(\theta_0)| < c\right) &= \Pr\left(-c < T_n(\theta_0) < c\right) \\ &= G_n(c) - G_n(-c) \\ &\equiv \overline{G}_n(c),\end{aligned}$$

which is a symmetric distribution function. The ideal critical value $c = q_n(\alpha)$ solves the equation

$$\overline{G}_n(q_n(\alpha)) = 1 - \alpha.$$

Equivalently, $q_n(\alpha)$ is the $1-\alpha$ quantile of the distribution of $|T_n(\theta_0)|$.

The bootstrap estimate is $q_n^*(\alpha)$, the $1-\alpha$ quantile of the distribution of $|T_n^*|$, or the number which solves the equation

$$\overline{G}_n^*(q_n^*(\alpha)) = G_n^*(q_n^*(\alpha)) - G_n^*(-q_n^*(\alpha)) = 1 - \alpha.$$

Computationally, $q_n^*(\alpha)$ is estimated from a bootstrap simulation by sorting the bootstrap t-statistics $|T_n^*| = \left|\hat{\theta}^* - \hat{\theta}\right| / s(\hat{\theta}^*)$, and taking the upper $\alpha\%$ quantile. The bootstrap test rejects if $|T_n(\theta_0)| > q_n^*(\alpha)$.

Let

$$C_4 = [\hat{\theta} - s(\hat{\theta})q_n^*(\alpha), \quad \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)],$$

where $q_n^*(\alpha)$ is the bootstrap critical value for a two-sided hypothesis test. C_4 is called the symmetric percentile-t interval. It is designed to work well since

$$\begin{aligned}\Pr\left(\theta_0 \in C_4\right) &= \Pr\left(\hat{\theta} - s(\hat{\theta})q_n^*(\alpha) \leq \theta_0 \leq \hat{\theta} + s(\hat{\theta})q_n^*(\alpha)\right) \\ &= \Pr\left(|T_n(\theta_0)| < q_n^*(\alpha)\right) \\ &\simeq \Pr\left(|T_n(\theta_0)| < q_n(\alpha)\right) \\ &= 1 - \alpha.\end{aligned}$$

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

$$W_n(\boldsymbol{\theta}) = n\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)' \hat{\mathbf{V}}_{\boldsymbol{\theta}}^{-1}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$$

or some other asymptotically chi-square statistic. Thus here $T_n(\boldsymbol{\theta}) = W_n(\boldsymbol{\theta})$. The ideal test rejects if $W_n \geq q_n(\alpha)$, where $q_n(\alpha)$ is the $(1-\alpha)\%$ quantile of the distribution of W_n. The bootstrap test rejects if $W_n \geq q_n^*(\alpha)$, where $q_n^*(\alpha)$ is the $(1-\alpha)\%$ quantile of the distribution of

$$W_n^* = n\left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right)' \hat{\mathbf{V}}_{\theta}^{*-1}\left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right).$$

Computationally, the critical value $q_n^*(\alpha)$ is found as the quantile from simulated values of W_n^*. Note in the simulation that the Wald statistic is a quadratic form in $\left(\hat{\boldsymbol{\theta}}^* - \hat{\boldsymbol{\theta}}\right)$, not $\left(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta}_0\right)$. [This is a typical mistake made by practitioners.]

10.8 Asymptotic Expansions

Let $T_n \in \mathbb{R}$ be a statistic such that

$$T_n \xrightarrow{d} \mathrm{N}(0, \sigma^2). \tag{10.3}$$

In some cases, such as when T_n is a t-ratio, then $\sigma^2 = 1$. In other cases σ^2 is unknown. Equivalently, writing $T_n \sim G_n(u, F)$ then for each u and F

$$\lim_{n\to\infty} G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right),$$

or

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + o(1). \tag{10.4}$$

While (10.4) 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*.

The following notation will be helpful. Let a_n be a sequence.

Definition 10.8.1 $a_n = o(1)$ *if* $a_n \to 0$ *as* $n \to \infty$

Definition 10.8.2 $a_n = O(1)$ *if* $|a_n|$ *is uniformly bounded.*

Definition 10.8.3 $a_n = o(n^{-r})$ *if* $n^r |a_n| \to 0$ *as* $n \to \infty$.

Basically, $a_n = O(n^{-r})$ if it declines to zero like n^{-r}.

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 10.8.1 *Under regularity conditions and (10.3),*

$$G_n(u, F) = \Phi\left(\frac{u}{\sigma}\right) + \frac{1}{n^{1/2}} g_1(u, F) + \frac{1}{n} g_2(u, F) + O(n^{-3/2})$$

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 10.8.1 is often called an **Edgeworth expansion**.

We can interpret Theorem 10.8.1 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).

[Side Note: When $T_n = \sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma$, a standardized sample mean, then

$$\begin{aligned} g_1(u) &= -\frac{1}{6}\kappa_3\left(u^2-1\right)\phi(u) \\ g_2(u) &= -\left(\frac{1}{24}\kappa_4\left(u^3-3u\right)+\frac{1}{72}\kappa_3^2\left(u^5-10u^3+15u\right)\right)\phi(u) \end{aligned}$$

where $\phi(u)$ is the standard normal pdf, and

$$\begin{aligned} \kappa_3 &= \mathbb{E}\left(X-\mu\right)^3/\sigma^3 \\ \kappa_4 &= \mathbb{E}\left(X-\mu\right)^4/\sigma^4 - 3 \end{aligned}$$

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.

10.9 One-Sided Tests

Using the expansion of Theorem 10.8.1, we can assess the accuracy of one-sided hypothesis tests and confidence regions based on an asymptotically normal t-ratio T_n. An asymptotic test is based on $\Phi(u)$.

To the second order, the exact distribution is

$$\Pr\left(T_n < u\right) = G_n(u, F_0) = \Phi(u) + \frac{1}{n^{1/2}} g_1(u, F_0) + O(n^{-1})$$

since $\sigma = 1$. The difference is

$$\begin{aligned} \Phi(u) - G_n(u, F_0) &= \frac{1}{n^{1/2}} g_1(u, F_0) + O(n^{-1}) \\ &= O(n^{-1/2}), \end{aligned}$$

so the order of the error is $O(n^{-1/2})$.

A bootstrap test is based on $G_n^*(u)$, which from Theorem 10.8.1 has the expansion

$$G_n^*(u) = G_n(u, F_n) = \Phi(u) + \frac{1}{n^{1/2}} g_1(u, F_n) + O(n^{-1}).$$

Because $\Phi(u)$ appears in both expansions, the difference between the bootstrap distribution and the true distribution is

$$G_n^*(u) - G_n(u, F_0) = \frac{1}{n^{1/2}}\left(g_1(u, F_n) - g_1(u, F_0)\right) + O(n^{-1}).$$

Since F_n converges to F at rate $\sqrt{n}$, and g_1 is continuous with respect to F, the difference $(g_1(u,F_n) - g_1(u,F_0))$ converges to 0 at rate $\sqrt{n}$. Heuristically,

$$\begin{aligned} g_1(u,F_n) - g_1(u,F_0) &\approx \frac{\partial}{\partial F} g_1(u,F_0)\,(F_n - F_0) \\ &= O(n^{-1/2}), \end{aligned}$$

The "derivative" $\frac{\partial}{\partial F} g_1(u,F)$ is only heuristic, as F is a function. We conclude that

$$G_n^*(u) - G_n(u,F_0) = O(n^{-1}),$$

or

$$\Pr\left(T_n^* \le u\right) = \Pr\left(T_n \le u\right) + O(n^{-1}),$$

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.

10.10 Symmetric Two-Sided Tests

If a random variable y has distribution function $H(u) = \Pr(y \le u)$, then the random variable $|y|$ has distribution function

$$\overline{H}(u) = H(u) - H(-u)$$

since

$$\begin{aligned} \Pr\left(|y| \le u\right) &= \Pr\left(-u \le y \le u\right) \\ &= \Pr\left(y \le u\right) - \Pr\left(y \le -u\right) \\ &= H(u) - H(-u). \end{aligned}$$

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_n has exact distribution $G_n(u,F)$, then $|T_n|$ 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 $|T_n|$. Since $T_n \xrightarrow{d} Z$, then $|T_n| \xrightarrow{d} |Z| \sim \overline{\Phi}$. Thus asymptotic critical values are taken from the $\overline{\Phi}$ distribution, and exact critical values are taken from the $\overline{G}_n(u,F_0)$ distribution. From Theorem 10.8.1, we can calculate that

$$\begin{aligned} \overline{G}_n(u,F) &= G_n(u,F) - G_n(-u,F) \\ &= \left(\Phi(u) + \frac{1}{n^{1/2}} g_1(u,F) + \frac{1}{n} g_2(u,F)\right) \\ &\quad - \left(\Phi(-u) + \frac{1}{n^{1/2}} g_1(-u,F) + \frac{1}{n} g_2(-u,F)\right) + O(n^{-3/2}) \\ &= \overline{\Phi}(u) + \frac{2}{n} g_2(u,F) + O(n^{-3/2}), \end{aligned} \tag{10.5}$$

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 (10.5) to the bootstrap distribution, we find

$$\overline{G}_n^*(u) = \overline{G}_n(u, F_n) = \overline{\Phi}(u) + \frac{2}{n} g_2(u, F_n) + O(n^{-3/2}).$$

Thus the difference between the bootstrap and exact distributions is

$$\begin{aligned} \overline{G}_n^*(u) - \overline{G}_n(u, F_0) &= \frac{2}{n}\left(g_2(u, F_n) - g_2(u, F_0)\right) + O(n^{-3/2}) \\ &= O(n^{-3/2}), \end{aligned}$$

the last equality because F_n converges to F_0 at rate $\sqrt{n}$, and g_2 is continuous in F. Another way of writing this is

$$\Pr\left(|T_n^*| < u\right) = \Pr\left(|T_n| < u\right) + O(n^{-3/2})$$

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.

10.11 Percentile Confidence Intervals

To evaluate the coverage rate of the percentile interval, set $T_n = \sqrt{n}\left(\hat{\theta} - \theta_0\right)$. We know that $T_n \xrightarrow{d} \mathrm{N}(0, V)$, which is not pivotal, as it depends on the unknown V. Theorem 10.8.1 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

$$G_n^*(u) = G_n(u, F_n) = \Phi\left(\frac{u}{\hat{\sigma}}\right) + O(n^{-1/2}),$$

where $\hat{\sigma} = V(F_n)$ is the bootstrap estimate of σ. The difference is

$$\begin{aligned} G_n^*(u) - G_n(u, F_0) &= \Phi\left(\frac{u}{\hat{\sigma}}\right) - \Phi\left(\frac{u}{\sigma}\right) + O(n^{-1/2}) \\ &= -\phi\left(\frac{u}{\sigma}\right)\frac{u}{\sigma}\left(\hat{\sigma} - \sigma\right) + O(n^{-1/2}) \\ &= O(n^{-1/2}) \end{aligned}$$

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.

10.12 Bootstrap Methods for Regression Models

The bootstrap methods we have discussed have set $G_n^*(u) = G_n(u, F_n)$, where F_n 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

$$\begin{aligned} y_i^* &= \boldsymbol{x}_i^{*\prime}\hat{\boldsymbol{\beta}} + e_i^* \\ \mathbb{E}\left(\boldsymbol{x}_i^* e_i^*\right) &= \boldsymbol{0} \end{aligned}$$

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 $y_i^* = \boldsymbol{x}_i^{*\prime}\hat{\boldsymbol{\beta}} + e_i^*$. There are different ways to impose independence. A non-parametric method is to sample the bootstrap errors e_i^* randomly from the OLS residuals $\{\hat{e}_1, ..., \hat{e}_n\}$. A parametric method is to generate the bootstrap errors e_i^* from a parametric distribution, such as the normal $e_i^* \sim \mathrm{N}(0, \hat{\sigma}^2)$.

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 $\mathbb{E}\left(e_i \mid \boldsymbol{x}_i\right) = 0$. 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

$$\begin{aligned} \mathbb{E}\left(e_i^* \mid \boldsymbol{x}_i\right) &= 0 \\ \mathbb{E}\left(e_i^{*2} \mid \boldsymbol{x}_i\right) &= \hat{e}_i^2 \\ \mathbb{E}\left(e_i^{*3} \mid \boldsymbol{x}_i\right) &= \hat{e}_i^3. \end{aligned}$$

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

$$\begin{aligned}\Pr\left(e_i^* = \left(\frac{1+\sqrt{5}}{2}\right)\hat{e}_i\right) &= \frac{\sqrt{5}-1}{2\sqrt{5}}\\ \Pr\left(e_i^* = \left(\frac{1-\sqrt{5}}{2}\right)\hat{e}_i\right) &= \frac{\sqrt{5}+1}{2\sqrt{5}}\end{aligned}$$

For each $\boldsymbol{x}_i$, you sample e_i^* using this two-point distribution.

Exercises

Exercise 10.1 Let $F_n(\boldsymbol{x})$ denote the EDF of a random sample. Show that

$$\sqrt{n}\left(F_n(\boldsymbol{x}) - F_0(\boldsymbol{x})\right) \xrightarrow{d} \mathrm{N}\left(0, F_0(\boldsymbol{x})\left(1 - F_0(\boldsymbol{x})\right)\right).$$

Exercise 10.2 Take a random sample $\{y_1, ..., y_n\}$ with $\mu = \mathbb{E}y_i$ and $\sigma^2 = \operatorname{var}(y_i)$. Let the statistic of interest be the sample mean $T_n = \overline{y}_n$. Find the population moments $\mathbb{E}T_n$ and $\operatorname{var}(T_n)$. Let $\{y_1^*, ..., y_n^*\}$ be a random sample from the empirical distribution function and let $T_n^* = \overline{y}_n^*$ be its sample mean. Find the bootstrap moments $\mathbb{E}T_n^*$ and $\operatorname{var}(T_n^*)$.

Exercise 10.3 Consider the following bootstrap procedure for a regression of y_i on $\boldsymbol{x}_i$. Let $\hat{\boldsymbol{\beta}}$ denote the OLS estimator from the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$, and $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ the OLS residuals.

(a) Draw a random vector $(\boldsymbol{x}^*, e^*)$ from the pair $\{(\boldsymbol{x}_i, \hat{e}_i) : i = 1, ..., n\}$. That is, draw a random integer i' from $[1, 2, ..., n]$, and set $\boldsymbol{x}^* = \boldsymbol{x}_{i'}$ and $e^* = \hat{e}_{i'}$. Set $y^* = \boldsymbol{x}^{*\prime}\hat{\boldsymbol{\beta}} + e^*$. 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 $\hat{\boldsymbol{\beta}}^*$ and any other statistic of interest.

Show that this bootstrap procedure is (numerically) *identical* to the non-parametric bootstrap.

Exercise 10.4 Consider the following bootstrap procedure. Using the non-parametric bootstrap, generate bootstrap samples, calculate the estimate $\hat{\theta}^*$ on these samples and then calculate

$$T_n^* = (\hat{\theta}^* - \hat{\theta})/s(\hat{\theta}),$$

where $s(\hat{\theta})$ is the standard error in the original data. Let $q_n^*(.05)$ and $q_n^*(.95)$ denote the 5% and 95% quantiles of T_n^*, and define the bootstrap confidence interval

$$C = \left[\hat{\theta} - s(\hat{\theta})q_n^*(.95), \quad \hat{\theta} - s(\hat{\theta})q_n^*(.05)\right].$$

Show that C exactly equals the Alternative percentile interval (not the percentile-t interval).

Exercise 10.5 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 $T_n = \hat{\theta}/s(\hat{\theta}) > c$ 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 $\hat{\theta}^*$ on these samples and then calculate

$$T_n^* = \hat{\theta}^*/s(\hat{\theta}^*).$$

Let $q_n^*(.95)$ denote the 95% quantile of T_n^*. You replace c with $q_n^*(.95)$, and thus reject $\mathbb{H}_0$ if $T_n = \hat{\theta}/s(\hat{\theta}) > q_n^*(.95)$. What is wrong with this procedure?

Exercise 10.6 Suppose that in an application, $\hat{\theta} = 1.2$ and $s(\hat{\theta}) = .2$. Using the non-parametric bootstrap, 1000 samples are generated from the bootstrap distribution, and $\hat{\theta}^*$ is calculated on each sample. The $\hat{\theta}^*$ are sorted, and the 2.5% and 97.5% quantiles of the $\hat{\theta}^*$ 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 10.7 The datafile `hprice1.dat` 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 11

NonParametric Regression

11.1 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.

11.2 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 $h > 0$. Later we will call h a **bandwidth**. This estimator can be written as

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} 1\left(|x_i - x| \leq h\right) y_i}{\sum_{i=1}^{n} 1\left(|x_i - x| \leq h\right)} \tag{11.1}$$

where $1(\cdot)$ is the indicator function. Alternatively, (11.1) can be written as

$$\widehat{m}(x) = \sum_{i=1}^{n} w_i(x) y_i \tag{11.2}$$

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 (11.2) is a weighted average of the y_i.

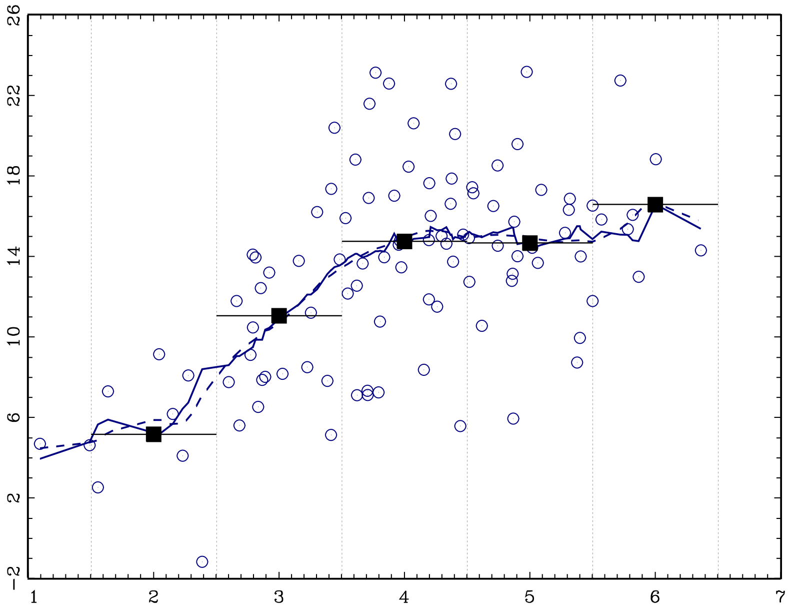


Figure 11.1: Scatter of (y_i, x_i) and Nadaraya-Watson regression

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 (11.1) is undefined for those values of x.

To visualize, Figure 11.1 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 (11.1) 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 11.1 by the first solid square. We repeat the calculation (11.1) 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 11.1 by the verticle dotted lines, and the estimates (11.1) 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 11.1 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 (11.1) there is no reason why we need to evaluate (11.1) only on a course grid. We can evaluate $\widehat{m}(x)$ for any set of values of x. In particular, we can evaluate (11.1) 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 11.1 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 11.1, 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 $m(x) = 10 \log(x)$.

11.3 Kernel Regression

One deficiency with the estimator (11.1) 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 11.1 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 (11.1) 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 (11.1) can be written as

$$\widehat{m}(x) = \frac{\sum_{i=1}^n k_0\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^n k_0\left(\frac{x_i - x}{h}\right)}. \tag{11.3}$$

The uniform density $k_0(u)$ is a special case of what is known as a **kernel function**.

Definition 11.3.1 *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 (11.1) is obtained by replacing the uniform kernel with any other kernel function:

$$\widehat{m}(x) = \frac{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right) y_i}{\sum_{i=1}^n k\left(\frac{x_i - x}{h}\right)}. \tag{11.4}$$

The estimator (11.4) also takes the form (11.2) 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 (11.4) is known as the **Nadaraya-Watson** estimator, the **kernel regression** estimator, or the **local constant** estimator.

The bandwidth h plays the same role in (11.4) as it does in (11.1). 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**

$$k_1(u) = \frac{3}{4}\left(1-u^2\right)1\left(|u|\le 1\right)$$

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 (11.4) 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 (11.4) using the Epanechnikov kernel and $h=1/2$ is also displayed in Figure 11.1 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

$$\sigma_k^2 = \int_{-\infty}^{\infty} u^2 k(u)du \tag{11.5}$$

$$R_k = \int_{-\infty}^{\infty} k(u)^2 du. \tag{11.6}$$

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	$k_0(u) = \frac{1}{2}1\left(\|u\|\le 1\right)$	$1/2$	$1/3$
Epanechnikov	$k_1(u) = \frac{3}{4}\left(1-u^2\right)1\left(\|u\|\le 1\right)$	$3/5$	$1/5$
Biweight	$k_2(u) = \frac{15}{16}\left(1-u^2\right)^2 1\left(\|u\|\le 1\right)$	$5/7$	$1/7$
Triweight	$k_3(u) = \frac{35}{32}\left(1-u^2\right)^3 1\left(\|u\|\le 1\right)$	$350/429$	$1/9$
Gaussian	$k_\phi(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)$	$1/\left(2\sqrt{\pi}\right)$	1

11.4 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, the local linear approximation 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

$$\left\{\widehat{\alpha}(x), \widehat{\beta}(x)\right\} = \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right)\left(y_i - \alpha - \beta\left(x_i - x\right)\right)^2.$$

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 11.2 displays the scatter plot of the same 100 observations from Figure 11.1, 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 11.2.

One interesting feature is that as $h \to \infty$, the LL estimator approaches the full-sample linear least-squares estimator $\widehat{m}(x) \to \hat{\alpha} + \hat{\beta}x$. 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.

11.5 Nonparametric Residuals and Regression Fit

The fitted regression at $x = x_i$ is $\widehat{m}(x_i)$ and the fitted residual is

$$\hat{e}_i = y_i - \widehat{m}(x_i).$$

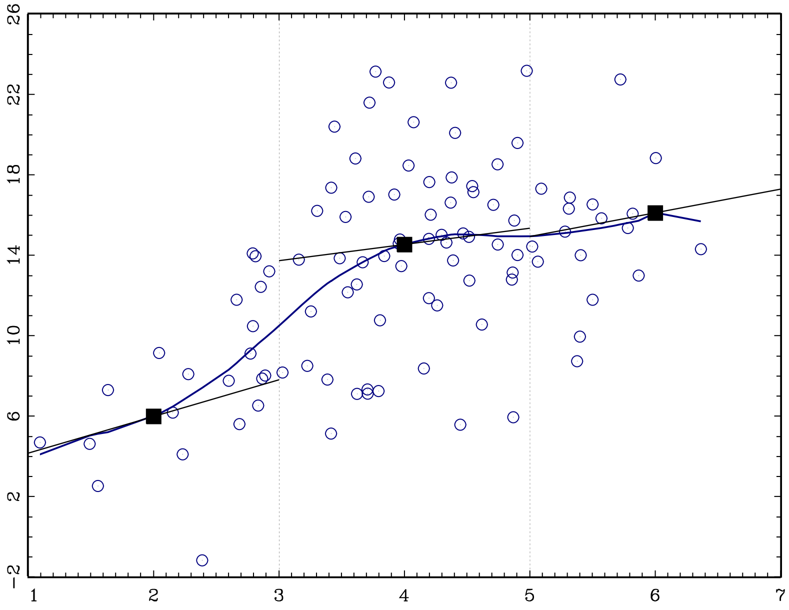


Figure 11.2: Scatter of (y_i, x_i) and Local Linear fitted regression

As a general rule, but especially when the bandwidth h is small, it is hard to view $\hat{e}_i$ as a good measure of the fit of the regression. As $h \to 0$ then $\widehat{m}(x_i) \to y_i$ and therefore $\hat{e}_i \to 0$. 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 $\hat{e}_i$ 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

$$\tilde{y}_i = \widetilde{m}_{-i}(x_i) = \frac{\sum_{j \neq i} k\left(\dfrac{x_j - x_i}{h}\right) y_j}{\sum_{j \neq i} k\left(\dfrac{x_j - x_i}{h}\right)}.$$

The leave-one-out residuals (or prediction errors) are the difference between the leave-one-out predicted values and the actual observation

$$\tilde{e}_i = y_i - \tilde{y}_i.$$

Since $\tilde{y}_i$ is not a function of y_i, there is no tendency for $\tilde{y}_i$ to overfit for small h. Consequently, $\tilde{e}_i$ is a good measure of the fit of the estimated nonparametric regression.

Similarly, the leave-one-out local-linear residual is $\tilde{e}_i = y_i - \widetilde{\alpha}_i$ 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

$$k_{ij} = k\left(\frac{x_j - x_i}{h}\right).$$

11.6 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

$$MSE_n(x,h) = \mathbb{E}\left(\widehat{m}(x,h) - m(x)\right)^2.$$

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

$$\begin{aligned} IMSE_n(h) &= \int MSE_n(x,h) f_x(x) dx \\ &= \int \mathbb{E}\left(\widehat{m}(x,h) - m(x)\right)^2 f_x(x) dx \end{aligned}$$

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 4.9. 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

$$\begin{aligned} MSFE_n(h) &= \mathbb{E}\left(y_{n+1} - \widehat{m}(x_{n+1},h)\right)^2 \\ &= \mathbb{E}\left(e_{n+1} + m(x_{n+1}) - \widehat{m}(x_{n+1},h)\right)^2 \\ &= \sigma^2 + \mathbb{E}\left(m(x_{n+1}) - \widehat{m}(x_{n+1},h)\right)^2 \\ &= \sigma^2 + \int \mathbb{E}\left(\widehat{m}(x,h) - m(x)\right)^2 f_x(x) dx \end{aligned}$$

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 4.9.1 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

$$\tilde{e}_i(h) = y_i - \widetilde{m}_{-i}(x_i,h)$$

where we are being explicit about the dependence on the bandwidth h. The mean squared leave-one-out residuals is

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \tilde{e}_i(h)^2.$$

This function of h is known as the **cross-validation criterion**.

The cross-validation bandwidth $\widehat{h}$ is the value which minimizes $CV(h)$

$$\widehat{h} = \underset{h \geq h_\ell}{\operatorname{argmin}}\, CV(h) \tag{11.7}$$

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 (11.7), 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 11.6.1

$$\mathbb{E}\left(CV(h)\right) = MSFE_{n-1}(h) = IMSE_{n-1}(h) + \sigma^2 \tag{11.8}$$

Theorem 11.6.1 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 11.3 displays the cross-validation criteria $CV(h)$ for the Nadaraya-Watson and Local Linear estimators using the data from Figure 11.1, 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 11.3 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 11.3, the NW estimator has a minimized CV criterion of 16.88, while the LL estimator has a 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 11.4 displays the fitted CEF estimates (NW and LL) using the bandwidths selected by cross-validation. Also displayed is the true CEF $m(x) = 10\ln(x)$. Notice that the nonparametric

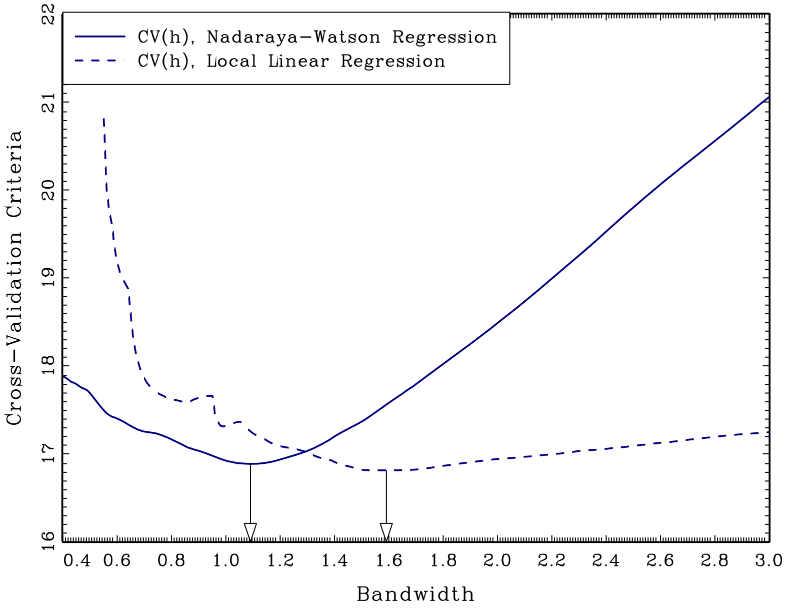


Figure 11.3: Cross-Validation Criteria, Nadaraya-Watson Regression and Local Linear Regression

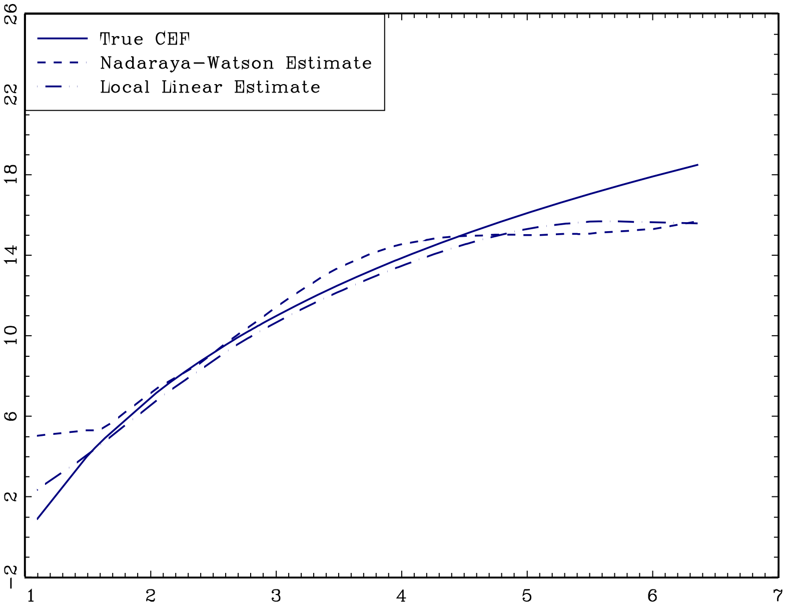


Figure 11.4: Nonparametric Estimates using data-dependent (CV) bandwidths

estimators with the CV-selected bandwidths (and especially the LL estimator) track the true CEF quite well.

Proof of Theorem 11.6.1. 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 $\tilde{e}_i(h) = m(x_i) - \widetilde{m}_{-i}(x_i, h) + e_i$, then

$$\begin{aligned}
\mathbb{E}\left(CV(h)\right) &= \mathbb{E}\left(\tilde{e}_i(h)^2\right) \\
&= \mathbb{E}\left(e_i^2\right) + \mathbb{E}\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right)^2 \\
&\quad + 2\mathbb{E}\left(\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right) e_i\right) \\
&= \sigma^2 + \mathbb{E}\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right)^2 . \qquad (11.9)
\end{aligned}$$

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

$$\mathbb{E}_{-i}\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right)^2 = \int \left(\widetilde{m}_{-i}(x, h) - m(x)\right)^2 f_x(x) dx.$$

Taking the unconditional expecation yields

$$\begin{aligned}
\mathbb{E}\left(\widetilde{m}_{-i}(x_i, h) - m(x_i)\right)^2 &= \mathbb{E}\int \left(\widetilde{m}_{-i}(x, h) - m(x)\right)^2 f_x(x) dx \\
&= IMSE_{n-1}(h)
\end{aligned}$$

where this is the IMSE of a sample of size $n-1$ as the estimator $\widetilde{m}_{-i}$ uses $n-1$ observations. Combined with (11.9) we obtain (11.8), as desired. ■

11.7 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 11.7.1 *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 $nh \to \infty$,*

$$\sqrt{nh}\left(\widehat{m}(x) - m(x) - h^2\sigma_k^2 B(x)\right) \xrightarrow{d} \mathrm{N}\left(0, \frac{R_k \sigma^2(x)}{f_x(x)}\right) \tag{11.10}$$

where σ_k^2 and R_k are defined in (11.5) and (11.6). 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 (11.10) 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 11.7.1 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 11.7.1 is that we can define the asymptotic MSE (AMSE) of $\widehat{m}(x)$ as the squared bias plus the asymptotic variance

$$AMSE\left(\widehat{m}(x)\right) = \left(h^2\sigma_k^2 B(x)\right)^2 + \frac{R_k\sigma^2(x)}{nhf_x(x)}. \tag{11.11}$$

Focusing on rates, this says

$$AMSE\left(\widehat{m}(x)\right) \sim h^4 + \frac{1}{nh} \tag{11.12}$$

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 (11.12). 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 11.7.2 *The bandwidth which minimizes the AMSE (11.12) 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 (11.10) 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 11.7.1 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 (11.10) is asymptotically negligible so Theorem 11.7.1 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 (11.10) 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.

11.8 Conditional Variance Estimation

Let's consider the problem of estimation of the conditional variance

$$\begin{aligned}\sigma^2(x) &= \operatorname{var}(y_i \mid x_i = x) \\ &= \mathbb{E}\left(e_i^2 \mid x_i = x\right).\end{aligned}$$

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 $\hat{e}_i = y_i - \widehat{m}(x_i)$ 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 $\tilde{e}_i = y_i - \widehat{m}_{-i}(x_i)$, as these are not subject to overfitting.

With this substitution the NW estimator of the conditional variance is

$$\tilde{\sigma}^2(x) = \frac{\sum_{i=1}^n k_i(x) \tilde{e}_i^2}{\sum_{i=1}^n k_i(x)}. \tag{11.13}$$

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 $\hat{\sigma}^2(x)$ 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 (11.13) 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 (11.13). They obtain the surprising result that the asymptotic distribution of this two-step estimator is identical to that of the one-step idealized estimator $\tilde{\sigma}^2(x)$.

11.9 Standard Errors

Theorem 11.7.1 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 (11.13). The density of x_i can be estimated using the methods from Section 20.1. Replacing these estimates into the formula for $V(x)$ we obtain the asymptotic variance estimate

$$\hat{V}(x) = \frac{R_k \hat{\sigma}^2(x)}{\hat{f}_x(x)}.$$

Then an asymptotic standard error for the kernel estimate $\widehat{m}(\boldsymbol{x})$ is

$$\hat{s}(x) = \sqrt{\frac{1}{nh}\hat{V}(x)}.$$

Plots of the estimated CEF $\widehat{m}(x)$ can be accompanied by confidence intervals $\widehat{m}(x) \pm 2\hat{s}(x)$. 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 $\widehat{m}(x) \pm 2\hat{s}(x)$ 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.

11.10 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 $\tilde{e}_i$ 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 11.10.1 *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 11.10.2 *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 12

Series Estimation

12.1 Approximation by Series

As we mentioned at the beginning of Chapter 11, 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

$$\begin{aligned} m_K(\boldsymbol{x}) &= \sum_{j=1}^{K} z_{jK}(\boldsymbol{x})\beta_{jK} \\ &= \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K \qquad (12.1) \end{aligned}$$

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 $\boldsymbol{x}$.

12.2 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, $t_2, \ldots, t_N$ 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 $\boldsymbol{x}_i$.

12.3 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 $\boldsymbol{\beta}_K = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_{2K}')'$.

12.4 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 estimation of 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.

12.5 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 $m(\boldsymbol{x})$:

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left|m_K(\boldsymbol{x}) - m(\boldsymbol{x})\right| \leq \varepsilon. \tag{12.2}$$

Thus the true unknown $m(\boldsymbol{x})$ can be arbitrarily well approximately by selecting a suitable polynomial.

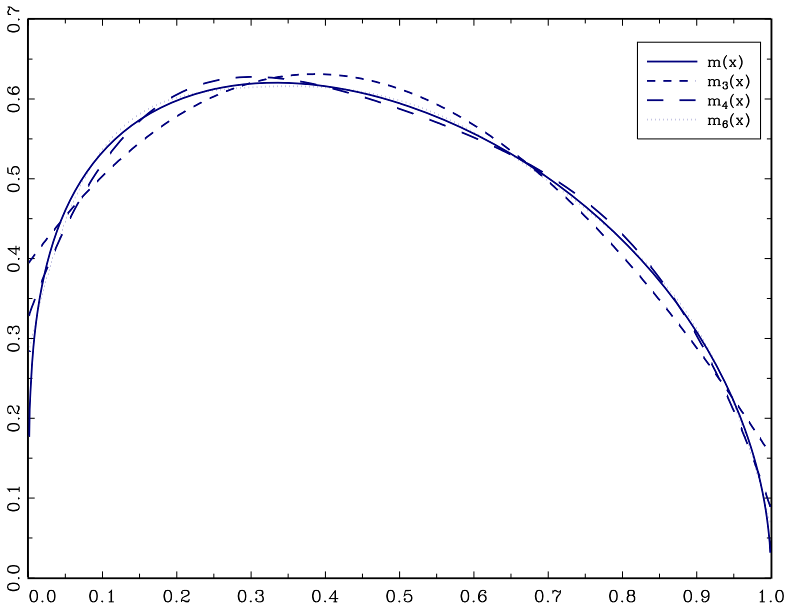


Figure 12.1: True CEF and Best Approximations

The result (12.2) can be stengthened. In particular, if the s^{th} derivative of $m(\boldsymbol{x})$ is continuous then the uniform approximation error satisfies

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |m_K(\boldsymbol{x}) - m(\boldsymbol{x})| = O\left(K^{-\alpha}\right) \tag{12.3}$$

as $K \to \infty$ where $\alpha = s/d$. This result is more useful than (12.2) because it gives a rate at which the approximation $m_K(\boldsymbol{x})$ approaches $m(\boldsymbol{x})$ as K increases.

Both (12.2) and (12.3) hold for spline approximations as well.

Intuitively, the number of derivatives s indexes the smoothness of the function $m(\boldsymbol{x})$. (12.3) 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 12.1 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 (12.1), then the coefficient of the best uniform approximation (12.3) is then

$$\boldsymbol{\beta}_K^* = \underset{\boldsymbol{\beta}_K}{\operatorname{argmin}} \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K - m(\boldsymbol{x})\right|. \tag{12.4}$$

The approximation error is

$$r_K^*(\boldsymbol{x}) = m(\boldsymbol{x}) - \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K^*.$$

We can write this as

$$m(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K^* + r_K^*(\boldsymbol{x}) \tag{12.5}$$

to emphasize that the true conditional mean can be written as the linear approximation plus error. A useful consequence of equation (12.3) is

$$\sup_{\boldsymbol{x}\in\mathcal{X}} |r_K^*(\boldsymbol{x})| \le O\left(K^{-\alpha}\right). \tag{12.6}$$

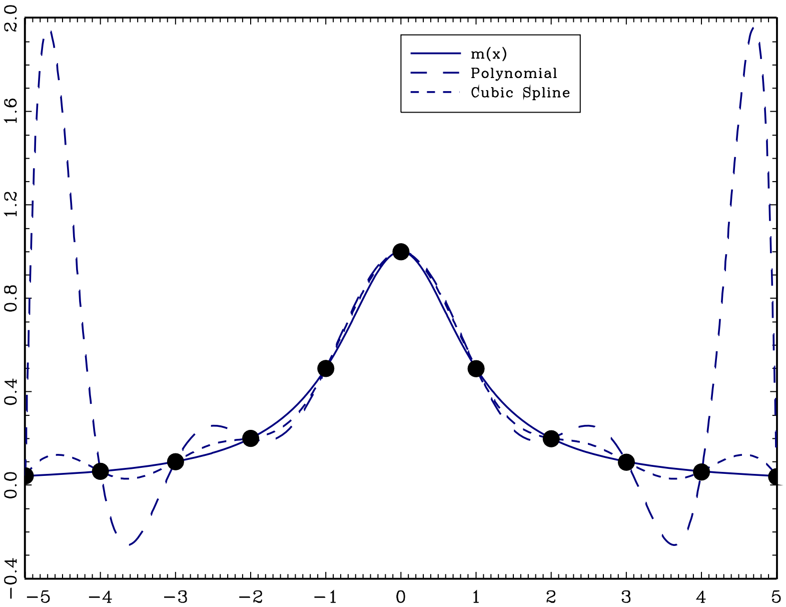


Figure 12.2: True CEF, polynomial interpolation, and spline interpolation

12.6 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 12.2. 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 12.2 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.

12.7 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

$$\widehat{m}_K(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})'\widehat{\boldsymbol{\beta}}_K. \tag{12.7}$$

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

$$r_K(\boldsymbol{x}) = m(\boldsymbol{x}) - \boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{\beta}_K. \tag{12.8}$$

The true CEF equation for y_i is

$$y_i = m(\boldsymbol{x}_i) + e_i \tag{12.9}$$

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

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{Z}_K\boldsymbol{\beta}_K + \boldsymbol{r}_K + \boldsymbol{e} \\ &= \boldsymbol{Z}_K\boldsymbol{\beta}_K + \boldsymbol{e}_K. \end{aligned}\tag{12.10}$$

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 $\boldsymbol{x}_i$

$$\zeta_K = \sup_{\boldsymbol{x}\in\mathcal{X}} \left(\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{Q}_K^{-1}\boldsymbol{z}_K(\boldsymbol{x})\right)^{1/2}. \tag{12.11}$$

$\boldsymbol{\zeta}_K$ 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 12.7.1

1. *For some $\alpha > 0$ the series approximation satisfies (12.3).*
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 12.7.1.1 through 12.7.1.3 concern properties of the regression model. Assumption 12.7.1.1 holds with $\alpha = s/d$ if $\mathcal{X}$ is compact and the s'th derivative of $m(\boldsymbol{x})$ is continuous. Assumption 12.7.1.2 allows for conditional heteroskedasticity, but requires the conditional variance to be bounded. Assumption 12.7.1.3 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 12.7.1.3 can be viewed as holding after transformation by an appropriate non-singular $\boldsymbol{B}_K$.

Assumption 12.7.1.4 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 12.7.1.4 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 12.5 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 (12.5) 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 (12.6). 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

$$\begin{aligned}r_K(\boldsymbol{x}) - r_K^*(\boldsymbol{x}) &= \boldsymbol{z}_K(\boldsymbol{x})'\left(\boldsymbol{\beta}_K^* - \boldsymbol{\beta}_K\right)\\ &= \boldsymbol{z}_K(\boldsymbol{x})'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}r_{Ki}^*\right).\end{aligned}\tag{12.12}$$

Observe that by the properties of projection

$$\mathbb{E}\left(\boldsymbol{r}_{Ki}^{*2}\right) - \mathbb{E}\left(\boldsymbol{r}_{Ki}^*\boldsymbol{z}_{Ki}\right)'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}r_{Ki}^*\right) \geq 0\tag{12.13}$$

and by (12.6)

$$\mathbb{E}\left(r_{Ki}^{*2}\right) = \int r_K^*(\boldsymbol{x})^2 f_x(\boldsymbol{x})d\boldsymbol{x} \leq O\left(K^{-2\alpha}\right).\tag{12.14}$$

Then applying the Schwarz inequality to (12.12), Definition (12.11), (12.13) and (12.14), we find

$$\begin{aligned}\left|r_K(\boldsymbol{x}) - r_K^*(\boldsymbol{x})\right| &\leq \left(\boldsymbol{z}_K(\boldsymbol{x})'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\boldsymbol{z}_K(\boldsymbol{x})\right)^{1/2}\\ &\quad\left(\mathbb{E}\left(r_{Ki}^*\boldsymbol{z}_{Ki}\right)'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}r_{Ki}^*\right)\right)^{1/2}\\ &\leq O\left(\zeta_K K^{-\alpha}\right).\end{aligned}\tag{12.15}$$

It follows that the best linear predictor approximation error satisfies

$$\sup_{\boldsymbol{x}\in\mathcal{X}}\left|r_K(\boldsymbol{x})\right| \leq O\left(\zeta_K K^{-\alpha}\right).\tag{12.16}$$

The bound (12.16) is probably not the best possible, but it shows that the best linear predictor satisfies a uniform approximation bound. Relative to (12.6), the rate is slower by the factor ζ_K. The bound (12.16) 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

$$\begin{aligned}\mathbb{E}r_{Ki}^2 &= \mathbb{E}\left(m(\boldsymbol{x}_i) - \boldsymbol{z}_{Ki}'\boldsymbol{\beta}_K\right)^2\\ &\leq \mathbb{E}\left(m(\boldsymbol{x}_i) - \boldsymbol{z}_{Ki}'\boldsymbol{\beta}_K^*\right)^2\\ &\leq O\left(K^{-2\alpha}\right)\end{aligned}\tag{12.17}$$

the final inequality by (12.14).

12.8 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

$$\hat{e}_{iK} = y_i - \widehat{m}_K(\boldsymbol{x}_i).$$

The leave-one-out prediction errors are

$$\begin{aligned}\tilde{e}_{iK} &= y_i - \widehat{m}_{K,-i}(\boldsymbol{x}_i) \\ &= y_i - \boldsymbol{z}_{Ki}'\widehat{\boldsymbol{\beta}}_{K,-i}\end{aligned}$$

where $\widehat{\boldsymbol{\beta}}_{K,-i}$ is the least-squares coefficient with the i'th observation omitted. Using (3.38) we can also write

$$\tilde{e}_{iK} = \hat{e}_{iK}(1 - h_{Kii})^{-1}$$

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 $\tilde{e}_{iK}$ are better estimates of the errors than the fitted residuals $\hat{e}_{iK}$, 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

$$\tilde{\sigma}_K^2 = \frac{1}{n}\sum_{i=1}^n \tilde{e}_{iK}^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_{iK}^2(1 - h_{Kii})^{-2}$$

and the prediction R^2 is

$$\widetilde{R}_K^2 = 1 - \frac{\sum_{i=1}^n \tilde{e}_{iK}^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

12.9 Cross-Validation Model Selection

The cross-validation criterion for selection of the number of series terms is the MSPE

$$CV(K) = \tilde{\sigma}_K^2 = \frac{1}{n}\sum_{i=1}^n \hat{e}_{iK}^2(1 - h_{Kii})^{-2}.$$

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 11.6.1, $CV(K)$ 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.

12.10 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 12.10.1 *Under Assumption 12.7.1, as $n \to \infty$,*

$$\left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)' \boldsymbol{Q}_K \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right) = O_p\left(\frac{K}{n}\right) + o_p\left(K^{-2\alpha}\right) \tag{12.18}$$

The proof of Theorem 12.10.1 is rather technical and deferred to Section 12.16.

The rate of convergence in (12.18) has two terms. The $O_p(K/n)$ term is due to estimation variance. Note in contrast that the corresponding rate would be $O_p(1/n)$ 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 (12.18) is due to the series approximation error.

Using Theorem 12.10.1 we can establish the following convergence rate for the estimated regression function.

Theorem 12.10.2 *Under Assumption 12.7.1, as $n \to \infty$,*

$$\int \left(\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})\right)^2 f_x(\boldsymbol{x}) d\boldsymbol{x} = O_p\left(\frac{K}{n}\right) + O_p\left(K^{-2\alpha}\right) \tag{12.19}$$

Theorem 12.10.2 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 12.10.2 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 (12.19) is $K = O\left(n^{1/(1+2\alpha)}\right)$, yielding an optimal rate of convergence in (12.19) 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 (12.19), using (12.7) and (12.8) we can write

$$\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x}) = \boldsymbol{z}_K(\boldsymbol{x})' \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right) - r_K(\boldsymbol{x}). \tag{12.20}$$

Since e_{Ki} are projection errors, they satisfy $\mathbb{E}\left(\boldsymbol{z}_{Ki}e_{Ki}\right) = 0$ and thus $\mathbb{E}\left(\boldsymbol{z}_{Ki}r_{Ki}\right) = 0$. 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 $\mathbb{E}r_{Ki}^2 = \int r_K(\boldsymbol{x})^2 f_x(\boldsymbol{x})d\boldsymbol{x}$. Then

$$\begin{aligned}
&\int \left(\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})\right)^2 f_x(\boldsymbol{x})d\boldsymbol{x} \\
&= \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)' \boldsymbol{Q}_K \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right) + \mathbb{E}r_{Ki}^2 \\
&\le O_p\left(\frac{K}{n}\right) + O_p\left(K^{-2\alpha}\right)
\end{aligned}$$

by (12.18) and (12.17), establishing (12.19).

12.11 Uniform Convergence

Theorem 12.10.2 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 12.11.1 *Under Assumption 12.7.1, then as $n \to \infty$,*

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})\right| = O_p\left(\sqrt{\frac{\zeta_K^2 K}{n}}\right) + O_p\left(\zeta_K K^{-\alpha}\right). \qquad (12.21)$$

Relative to Theorem 12.10.2, 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 (12.21) notice that the first term is $o_p(1)$ under Assumption 12.7.1.4. 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 (12.21) is straightforward to show using (12.18). Using (12.20), the Triangle Inequality, the Schwarz inequality (A.15), Definition (12.11), (12.18) and (12.16),

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in\mathcal{X}} \left|\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})\right| \\
&\le \sup_{\boldsymbol{x}\in\mathcal{X}} \left|\boldsymbol{z}_K(\boldsymbol{x})'\left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)\right| + \sup_{\boldsymbol{x}\in\mathcal{X}} \left|r_K(\boldsymbol{x})\right| \\
&\le \sup_{\boldsymbol{x}\in\mathcal{X}} \left(\boldsymbol{z}_K(\boldsymbol{x})'\boldsymbol{Q}_K^{-1}\boldsymbol{z}_K(\boldsymbol{x})\right)^{1/2} \left(\left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)' \boldsymbol{Q}_K \left(\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K\right)\right)^{1/2} \\
&+ O\left(\zeta_K K^{-\alpha}\right) \\
&\le \zeta_K \left(O_p\left(\frac{K}{n}\right) + O_p\left(K^{-2\alpha}\right)\right)^{1/2} + O\left(\zeta_K K^{-\alpha}\right), \\
&= O_p\left(\sqrt{\frac{\zeta_K^2 K}{n}}\right) + O_p\left(\zeta_K K^{-\alpha}\right). \qquad (12.22)
\end{aligned}$$

This is (12.21).

12.12 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

$$\hat{\theta}_K = a(\widehat{m}_K) = \boldsymbol{a}_K'\widehat{\boldsymbol{\beta}}_K$$

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 $\hat{\theta}_K$ 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 $\hat{\theta}_K$ is still asymptotically normal, and v_K is still the appropriate variance for $\hat{\theta}_K$. 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 12.12.1 *Under Assumption 12.7.1, 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* $n \to \infty$,

$$\frac{\sqrt{n}\left(\hat{\theta}_K - \theta + a(r_K)\right)}{v_K^{1/2}} \xrightarrow{d} \mathrm{N}(0,1) \qquad (12.23)$$

The proof of Theorem 12.12.1 can be found in Section 12.16.

Theorem 12.12.1 shows that the estimator $\hat{\theta}_K$ 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 12.12.1 requires $\zeta_K K^{-\alpha} = O(1)$, which is similar to that found in Theorem 12.11.1 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 12.12.1 includes parametric least-squares as a special case with explicit attention to estimation bias.

One useful message from Theorem 12.12.1 is that the classic variance formula v_K for $\hat{\theta}_K$ still applies for series regression. Indeed, we can estimate the asymptotic variance using the standard White formula

$$\hat{v}_K = \boldsymbol{a}_K' \widehat{\boldsymbol{Q}}_K^{-1} \widehat{\boldsymbol{\Omega}}_K \widehat{\boldsymbol{Q}}_K^{-1} \boldsymbol{a}_K$$

$$\widehat{\boldsymbol{\Omega}}_K = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\hat{e}_{iK}^2$$

$$\widehat{\boldsymbol{Q}}_K = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'.$$

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 $\hat{v}_K / v_K \xrightarrow{p} 1$ as $n \to \infty$ and thus the distribution in (12.23) is unchanged if v_K is replaced with $\hat{v}_K$.

Theorem 12.12.1 shows that the estimator $\hat{\theta}_K$ 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 $\hat{\theta}_K$ for θ shown in Theorem 12.12.1 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.

12.13 Asymptotic Normality with Undersmoothing

An unpleasant aspect about Theorem 12.12.1 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 12.13.1 *Under Assumption 12.7.1, 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*

$$\frac{\sqrt{n}\left(\hat{\theta}_K - \theta\right)}{v_K^{1/2}} \xrightarrow{d} \mathrm{N}\left(0,1\right). \tag{12.24}$$

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 (12.6).

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 12.13.1, as the distribution (12.24) 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 (12.24). 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 (12.24) 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 (12.23) and (12.24) 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 (12.23) is more informative than (12.24). It shows that the asymptotic distribution is normal but has a non-trivial finite sample bias.

12.14 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 12.12.1 and 12.13.1 for this estimator.

Theorem 12.14.1 *Under Assumption 12.7.1, 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* $n \to \infty$,

$$\frac{\sqrt{n}\left(\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x}) + \boldsymbol{r}_K(\boldsymbol{x})\right)}{v_K^{1/2}(\boldsymbol{x})} \xrightarrow{d} \mathrm{N}\left(0,1\right) \tag{12.25}$$

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*

$$\frac{\sqrt{n}\left(\widehat{m}_K(\boldsymbol{x}) - m(\boldsymbol{x})\right)}{v_K^{1/2}(\boldsymbol{x})} \xrightarrow{d} \mathrm{N}\left(0,1\right) \tag{12.26}$$

There are two important features about the asymptotic distribution (12.25).

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, (12.25) 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 (12.26) 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.

12.15 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.

12.16 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 12.10.1, 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 12.16.1 *Under Assumption 12.7.1 and $\boldsymbol{Q}_K = \boldsymbol{I}_K$, as $n \to \infty$,*

$$\left\|\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K\right\| = o_p(1) \tag{12.27}$$

and

$$\lambda_{\min}(\widehat{\boldsymbol{Q}}_K) \xrightarrow{p} 1. \tag{12.28}$$

Proof. Since

$$\left\|\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K\right\|^2 = \sum_{j=1}^{K}\sum_{\ell=1}^{K}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{z}_{jKi}\boldsymbol{z}_{\ell Ki} - \mathbb{E}\boldsymbol{z}_{jKi}\boldsymbol{z}_{\ell Ki}\right)\right)^2$$

then

$$\begin{aligned}\mathbb{E}\left\|\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K\right\|^2 &= \sum_{j=1}^{K}\sum_{\ell=1}^{K}\operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_{jKi}\boldsymbol{z}_{\ell Ki}\right)\\ &= n^{-1}\sum_{j=1}^{K}\sum_{\ell=1}^{K}\operatorname{var}\left(\boldsymbol{z}_{jKi}\boldsymbol{z}_{\ell Ki}\right)\\ &\le n^{-1}\mathbb{E}\sum_{j=1}^{K}\boldsymbol{z}_{jKi}^2\sum_{\ell=1}^{K}\boldsymbol{z}_{\ell Ki}^2\\ &= n^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki}\right)^2. \end{aligned} \tag{12.29}$$

Since $\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki} \le \zeta_K^2$ by definition (12.11) and using (A.1) we find

$$\mathbb{E}\left(\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki}\right) = \operatorname{tr}\left(\mathbb{E}\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right) = \operatorname{tr}\boldsymbol{I}_K = K, \tag{12.30}$$

so that

$$\mathbb{E}\left(\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki}\right)^2 \le \zeta_K^2 K \tag{12.31}$$

and hence (12.29) is $o(1)$ under Assumption 12.7.1.4. Theorem 5.11.1 shows that this implies (12.27).

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

$$\left|\lambda_{\min}(\widehat{\boldsymbol{Q}}_K) - 1\right| = \left|\lambda_{\min}(\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K)\right| \le \left(\sum_{\ell=1}^{K}\lambda_\ell^2\right)^{1/2} = \left\|\widehat{\boldsymbol{Q}}_K - \boldsymbol{I}_K\right\|$$

where the second equality is (A.17). This is $o_p(1)$ by (12.27), establishing (12.28). ■

Proof of Theorem 12.10.1. As above, assume that the regressors have been transformed so that $\boldsymbol{Q}_K = \boldsymbol{I}_K$.

From expression (12.10) we can substitute to find

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_K - \boldsymbol{\beta}_K &= \left(\boldsymbol{Z}_K'\boldsymbol{Z}_K\right)^{-1}\boldsymbol{Z}_K'\boldsymbol{e}_K.\\ &= \widehat{\boldsymbol{Q}}_K^{-1}\left(\frac{1}{n}\boldsymbol{Z}_K'\boldsymbol{e}_K\right)\end{aligned} \tag{12.32}$$

Using (12.32) and the Quadratic Inequality (A.23),

$$\begin{aligned}&\left(\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K\right)'\left(\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K\right)\\ &=n^{-2}\left(\boldsymbol{e}_K'\boldsymbol{Z}_K\right)\widehat{\boldsymbol{Q}}_K^{-1}\widehat{\boldsymbol{Q}}_K^{-1}\left(\boldsymbol{Z}_K'\boldsymbol{e}_K\right)\\ &\le\left(\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}\right)\right)^2 n^{-2}\left(\boldsymbol{e}_K'\boldsymbol{Z}_K\boldsymbol{Z}_K'\boldsymbol{e}_K\right).\end{aligned}\tag{12.33}$$

Observe that (12.28) implies

$$\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}\right)=\left(\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K\right)\right)^{-1}=O_p(1).\tag{12.34}$$

Since $e_{Ki}=e_i+r_{Ki}$, and using Assumption 12.7.1.2 and (12.16), then

$$\sup_i\mathbb{E}\left(e_{Ki}^2|\boldsymbol{x}_i\right)=\overline{\sigma}^2+\sup_i r_{Ki}^2\le\overline{\sigma}^2+O\left(\zeta_K^2K^{-2\alpha}\right).\tag{12.35}$$

As e_{Ki} are projection errors, they satisfy $\mathbb{E}\left(\boldsymbol{z}_{Ki}e_{Ki}\right)=0$. Since the observations are independent, using (12.30) and (12.35), then

$$\begin{aligned}n^{-2}\mathbb{E}\left(\boldsymbol{e}_K'\boldsymbol{Z}_K\boldsymbol{Z}_K'\boldsymbol{e}_K\right)&=n^{-2}\mathbb{E}\left(\sum_{i=1}^n e_{Ki}\boldsymbol{z}_{Ki}'\sum_{ij=1}^n\boldsymbol{z}_{Kj}e_{Kj}\right)\\ &=n^{-2}\sum_{i=1}^n\mathbb{E}\left(\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki}e_{Ki}^2\right)\\ &\le n^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki}\right)\sup_i\mathbb{E}\left(e_{Ki}^2|\boldsymbol{x}_i\right)\\ &\le\overline{\sigma}^2\frac{K}{n}+O\left(\frac{\zeta_K^2K^{1-2\alpha}}{n}\right)\\ &=\overline{\sigma}^2\frac{K}{n}+o\left(K^{-2\alpha}\right)\end{aligned}\tag{12.36}$$

since $\zeta_K^2K/n=o(1)$ by Assumption 12.7.1.4. Theorem 5.11.1 shows that this implies

$$n^{-2}\boldsymbol{e}_K'\boldsymbol{Z}_K\boldsymbol{Z}_K'\boldsymbol{e}_K=O_p\left(n^{-2}\right)+o_p\left(K^{-2\alpha}\right).\tag{12.37}$$

Together, (12.33), (12.34) and (12.37) imply (12.18). ■

Proof of Theorem 12.12.1. 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 (12.32) we find

$$\begin{aligned}\hat{\theta}_K-\theta+a\left(r_K\right)&=\boldsymbol{a}_K'\left(\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K\right)\\ &=\frac{1}{n}\boldsymbol{a}_K'\widehat{\boldsymbol{Q}}_K^{-1}\boldsymbol{Z}_K'\boldsymbol{e}_K\end{aligned}$$

and thus

$$\begin{aligned}
\sqrt{\frac{n}{v_k}}\left(\hat{\theta}_K-\theta_K+a\left(r_K\right)\right) &= \sqrt{\frac{n}{v_k}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{\beta}}_K-\boldsymbol{\beta}_K\right) \\
&= \sqrt{\frac{1}{nv_k}}\boldsymbol{a}_K'\widehat{\boldsymbol{Q}}_K^{-1}\boldsymbol{Z}_K'\boldsymbol{e}_K \\
&= \frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\boldsymbol{Z}_K'\boldsymbol{e}_K \qquad (12.38)\\
&+ \frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{e} \qquad (12.39)\\
&+ \frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{r}_K. \qquad (12.40)
\end{aligned}$$

where we have used $\boldsymbol{e}_K=\boldsymbol{e}+\boldsymbol{r}_K$. We now take the terms in (12.38)-(12.40) separately.

First, take (12.38). We can write

$$\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\boldsymbol{Z}_K'\boldsymbol{e}_K=\frac{1}{\sqrt{nv_K}}\sum_{i=1}^n\boldsymbol{a}_K'\boldsymbol{z}_{Ki}e_{Ki}. \qquad (12.41)$$

Observe that $\boldsymbol{a}_K'\boldsymbol{z}_{Ki}e_{Ki}$ are independent across i, mean zero, and have variance

$$\mathbb{E}\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}e_{Ki}\right)^2=\boldsymbol{a}_K'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'e_{Ki}^2\right)\boldsymbol{a}_K=v_K.$$

We will apply the Lindeberg CLT 5.7.2, for which it is sufficient to verify Lyapunov's condition (5.6):

$$\frac{1}{n^2v_K^2}\sum_{i=1}^n\mathbb{E}\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}e_{Ki}\right)^4=\frac{1}{nv_K^2}\mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4e_{Ki}^4\right)\to 0. \qquad (12.42)$$

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 $\mathbb{E}\left(e_i^4|\boldsymbol{x}_i\right)\le\kappa$

$$\sup_i\mathbb{E}\left(e_{Ki}^4|\boldsymbol{x}_i\right)\le 8\sup_i\left(\mathbb{E}\left(e_i^4|\boldsymbol{x}_i\right)+r_{Ki}^4\right)\le 8\left(\kappa+\kappa_1\right). \qquad (12.43)$$

Using (12.43), the Schwarz Inequality, and (12.31)

$$\begin{aligned}
\mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4e_{Ki}^4\right) &= \mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4\mathbb{E}\left(e_{Ki}^4|\boldsymbol{x}_i\right)\right) \\
&\le 8\left(\kappa+\kappa_1\right)\mathbb{E}\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4 \\
&\le 8\left(\kappa+\kappa_1\right)\left(\boldsymbol{a}_K'\boldsymbol{a}_K\right)^2\mathbb{E}\left(\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki}\right)^2 \\
&= 8\left(\kappa+\kappa_1\right)\left(\boldsymbol{a}_K'\boldsymbol{a}_K\right)^2\zeta_K^2K. \qquad (12.44)
\end{aligned}$$

Since $\mathbb{E}\left(e_{Ki}^2|\boldsymbol{x}_i\right)=\mathbb{E}\left(e_i^2|\boldsymbol{x}_i\right)+r_{Ki}^2\ge\underline{\sigma}^2$,

$$\begin{aligned}
v_K &= \boldsymbol{a}_K'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'e_{Ki}^2\right)\boldsymbol{a}_K \\
&\ge \underline{\sigma}^2\boldsymbol{a}_K'\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)\boldsymbol{a}_K \\
&= \underline{\sigma}^2\boldsymbol{a}_K'\boldsymbol{a}_K. \qquad (12.45)
\end{aligned}$$

Equation (12.44) and (12.45) combine to show that

$$\frac{1}{nv_K^2}\mathbb{E}\left(\left(\boldsymbol{a}_K'\boldsymbol{z}_{Ki}\right)^4e_{Ki}^4\right)\le\frac{8\left(\kappa+\kappa_1\right)}{\underline{\sigma}^4}\frac{\zeta_K^2K}{n}=o(1)$$

under Assumption 12.7.1.4. This establishes Lyapunov's condition (12.42). Hence the Lindeberg CLT applies to (12.41) and we conclude

$$\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\boldsymbol{Z}_K'\boldsymbol{e}_K \xrightarrow{d} \mathrm{N}(0,1). \tag{12.46}$$

Second, take (12.39). Since $\mathbb{E}(\boldsymbol{e}\mid\boldsymbol{X})=0$, then applying $\mathbb{E}\left(e_i^2|\boldsymbol{x}_i\right)\leq\bar{\sigma}^2$, the Schwarz and Norm Inequalities, (12.45), (12.34) and (12.27),

$$\begin{aligned}
&\mathbb{E}\left(\left(\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{e}\right)^2\mid\boldsymbol{X}\right)\\
&=\frac{1}{nv_K}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\mathbb{E}\left(\boldsymbol{e}\boldsymbol{e}'\mid\boldsymbol{X}\right)\boldsymbol{Z}_K\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{a}_K\\
&\leq\frac{\bar{\sigma}^2}{v_K}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\widehat{\boldsymbol{Q}}_K\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{a}_K\\
&=\frac{\bar{\sigma}^2}{v_K}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K-\boldsymbol{I}_K\right)\widehat{\boldsymbol{Q}}_K^{-1}\left(\widehat{\boldsymbol{Q}}_K-\boldsymbol{I}_K\right)\boldsymbol{a}_K\\
&\leq\frac{\bar{\sigma}^2\boldsymbol{a}_K'\boldsymbol{a}_K}{v_K}\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}\right)\left\|\widehat{\boldsymbol{Q}}_K-\boldsymbol{I}_K\right\|^2\\
&\leq\frac{\bar{\sigma}^2}{\underline{\sigma}^2}o_p(1).
\end{aligned}$$

This establishes

$$\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{e}\xrightarrow{p}0. \tag{12.47}$$

Third, take (12.40). By the Cauchy-Schwarz inequality, (12.45), and the Quadratic Inequality,

$$\begin{aligned}
&\left(\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{r}_K\right)^2\\
&\leq\frac{\boldsymbol{a}_K'\boldsymbol{a}_K}{nv_K}\boldsymbol{r}_K'\boldsymbol{Z}_K\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{r}_K\\
&\leq\frac{\boldsymbol{1}}{\underline{\sigma}^2}\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)^2\frac{1}{n}\boldsymbol{r}_K'\boldsymbol{Z}_K\boldsymbol{Z}_K'\boldsymbol{r}_K.
\end{aligned} \tag{12.48}$$

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 (12.17)

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n}\boldsymbol{r}_K'\boldsymbol{Z}_K\boldsymbol{Z}_K'\boldsymbol{r}_K\right)&=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n r_{Ki}\boldsymbol{z}_{Ki}'\sum_{ij=1}^n\boldsymbol{z}_{Kj}r_{Kj}\right)\\
&=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}_{Ki}'\boldsymbol{z}_{Ki}r_{Ki}^2\right)\\
&\leq\zeta_K^2\mathbb{E}\left(r_{Ki}^2\right)\\
&=O\left(\zeta_K^2K^{-2\alpha}\right)\\
&=O(1)
\end{aligned}$$

since $\zeta_KK^{-2}=O(1)$. Thus $\frac{1}{n}\boldsymbol{r}_K'\boldsymbol{Z}_K\boldsymbol{Z}_K'\boldsymbol{r}_K=O_p(1)$. This means that (12.48) is $o_p(1)$ since (12.28) implies

$$\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)=\lambda_{\max}\left(\widehat{\boldsymbol{Q}}_K^{-1}\right)-1=o_p(1). \tag{12.49}$$

Equivalently,

$$\frac{1}{\sqrt{nv_K}}\boldsymbol{a}_K'\left(\widehat{\boldsymbol{Q}}_K^{-1}-\boldsymbol{I}_K\right)\boldsymbol{Z}_K'\boldsymbol{r}_K\xrightarrow{p}0. \tag{12.50}$$

Equations (12.46), (12.47) and (12.50) applied to (12.38)-(12.40) show that

$$\sqrt{\frac{n}{v_k}}\left(\hat{\theta}_K-\theta_K+a\left(r_K\right)\right)\xrightarrow{d}\mathrm{N}\left(0,1\right)$$

completing the proof. ■

Proof of Theorem 12.13.1. 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 12.12.1 are satisfied. It is thus sufficient to show that

$$\sqrt{\frac{n}{v_k}}a\left(r_K\right)=o(1).$$

From (12.12)

$$\begin{aligned}r_K(\boldsymbol{x})&=r_K^*(\boldsymbol{x})+\boldsymbol{z}_K(\boldsymbol{x})'\gamma_K\\ \gamma_K&=\mathbb{E}\left(\boldsymbol{z}_{Ki}\boldsymbol{z}_{Ki}'\right)^{-1}\mathbb{E}\left(\boldsymbol{z}_{Ki}r_{Ki}^*\right).\end{aligned}$$

Thus by linearity, applying (12.45), and the Schwarz inequality

$$\begin{aligned}\sqrt{\frac{n}{v_k}}a\left(r_K\right)&=\sqrt{\frac{n}{v_k}}\left(a\left(r_K^*\right)+\boldsymbol{a}_K'\gamma_K\right)\\ &\leq\frac{n^{1/2}}{\underline{\sigma}^2\left(\boldsymbol{a}_K'\boldsymbol{a}_K\right)^{1/2}}a\left(r_K^*\right) &(12.51)\\ &+\frac{\left(n\gamma_K'\gamma_K\right)^{1/2}}{\underline{\sigma}}. &(12.52)\end{aligned}$$

By assumption, $n^{1/2}a\left(r_K^*\right)=O\left(n^{1/2}K^{-\alpha}\right)=o(1)$. By (12.14) 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 (12.51) and (12.52) are $o(1)$, as required. ■

Chapter 13

Generalized Method of Moments

13.1 Overidentified Linear Model

Consider the linear model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + \boldsymbol{x}_{2i}'\boldsymbol{\beta}_2 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0} \end{aligned}$$

where $\boldsymbol{x}_{1i}$ is $k \times 1$ and $\boldsymbol{x}_{2i}$ is $r \times 1$ with $\ell = k + r$. We know that without further restrictions, an asymptotically efficient estimator of $\boldsymbol{\beta}$ is the OLS estimator. Now suppose that we are given the information that $\boldsymbol{\beta}_2 = \mathbf{0}$. Now we can write the model as

$$\begin{aligned} y_i &= \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1 + e_i \\ \mathbb{E}\left(\boldsymbol{x}_i e_i\right) &= \mathbf{0}. \end{aligned}$$

In this case, how should $\boldsymbol{\beta}_1$ be estimated? One method is OLS regression of y_i on $\boldsymbol{x}_{1i}$ alone. This method, however, is not necessarily efficient, as there are ℓ restrictions in $\mathbb{E}\left(\boldsymbol{x}_i e_i\right) = \mathbf{0}$, while $\boldsymbol{\beta}_1$ is of dimension $k < \ell$. This situation is called **overidentified**. There are $\ell - k = r$ more moment restrictions than free parameters. We call r the **number of overidentifying restrictions**.

This is a special case of a more general class of moment condition models. Let $\boldsymbol{g}(y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta})$ be an $\ell \times 1$ function of a $k \times 1$ parameter $\boldsymbol{\beta}$ with $\ell \geq k$ such that

$$\mathbb{E}\boldsymbol{g}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\beta}_0) = \mathbf{0} \tag{13.1}$$

where $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. In our previous example, $\boldsymbol{g}(y, \boldsymbol{z}, \boldsymbol{\beta}) = \boldsymbol{z}\cdot(y - \boldsymbol{x}_1'\boldsymbol{\beta}_1)$. In econometrics, this class of models are called **moment condition models**. In the statistics literature, these are known as **estimating equations**.

As an important special case we will devote special attention to linear moment condition models, which can be written as

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= \mathbf{0}. \end{aligned}$$

where the dimensions of $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are $k \times 1$ and $\ell \times 1$, with $\ell \geq k$. If $k = \ell$ the model is **just identified**, otherwise it is **overidentified**. The variables $\boldsymbol{x}_i$ may be components and functions of $\boldsymbol{z}_i$, but this is not required. This model falls in the class (13.1) by setting

$$\boldsymbol{g}(y, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\beta}_0) = \boldsymbol{z}\cdot(y - \boldsymbol{x}'\boldsymbol{\beta}) \tag{13.2}$$

13.2 GMM Estimator

Define the sample analog of (13.2)

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_i \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right) = \frac{1}{n}\left(\boldsymbol{Z}'\boldsymbol{y} - \boldsymbol{Z}'\boldsymbol{X}\boldsymbol{\beta}\right). \tag{13.3}$$

The method of moments estimator for $\boldsymbol{\beta}$ is defined as the parameter value which sets $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \boldsymbol{0}$. This is generally not possible when $\ell > k$, as there are more equations than free parameters. The idea of the generalized method of moments (GMM) is to define an estimator which sets $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$ "close" to zero.

For some $\ell \times \ell$ weight matrix $\boldsymbol{W}_n > 0$, let

$$J_n(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \boldsymbol{W}_n \, \overline{\boldsymbol{g}}_n(\boldsymbol{\beta}).$$

This is a non-negative measure of the "length" of the vector $\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$. For example, if $\boldsymbol{W}_n = \boldsymbol{I}$, then, $J_n(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})'\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = n \cdot \left\|\overline{\boldsymbol{g}}_n(\boldsymbol{\beta})\right\|^2$, the square of the Euclidean length. The GMM estimator minimizes $J_n(\boldsymbol{\beta})$.

Definition 13.2.1 $\widehat{\boldsymbol{\beta}}_{GMM} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, J_n(\boldsymbol{\beta})$.

Note that if $k = \ell$, then $\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}$, and the GMM estimator is the method of moments estimator. The first order conditions for the GMM estimator are

$$\begin{aligned}\boldsymbol{0} &= \frac{\partial}{\partial \boldsymbol{\beta}} J_n(\widehat{\boldsymbol{\beta}}) \\ &= 2\frac{\partial}{\partial \boldsymbol{\beta}} \overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})' \boldsymbol{W}_n \overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}) \\ &= -2\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right) \boldsymbol{W}_n \left(\frac{1}{n}\boldsymbol{Z}'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\right)\end{aligned}$$

so

$$2\left(\boldsymbol{X}'\boldsymbol{Z}\right) \boldsymbol{W}_n \left(\boldsymbol{Z}'\boldsymbol{X}\right) \widehat{\boldsymbol{\beta}} = 2\left(\boldsymbol{X}'\boldsymbol{Z}\right) \boldsymbol{W}_n \left(\boldsymbol{Z}'\boldsymbol{y}\right)$$

which establishes the following.

Proposition 13.2.1

$$\widehat{\boldsymbol{\beta}}_{GMM} = \left(\left(\boldsymbol{X}'\boldsymbol{Z}\right) \boldsymbol{W}_n \left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1} \left(\boldsymbol{X}'\boldsymbol{Z}\right) \boldsymbol{W}_n \left(\boldsymbol{Z}'\boldsymbol{y}\right).$$

While the estimator depends on $\boldsymbol{W}_n$, the dependence is only up to scale, for if $\boldsymbol{W}_n$ is replaced by $c\boldsymbol{W}_n$ for some $c > 0$, $\widehat{\boldsymbol{\beta}}_{GMM}$ does not change.

13.3 Distribution of GMM Estimator

Assume that $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{W} > 0$. 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

$$\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{W}_n\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right) \xrightarrow{p} \boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}$$

and

$$\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\boldsymbol{W}_n\left(\frac{1}{\sqrt{n}}\boldsymbol{Z}'\boldsymbol{e}\right) \xrightarrow{d} \boldsymbol{Q}'\boldsymbol{W}\mathrm{N}\left(\boldsymbol{0}, \boldsymbol{\Omega}\right).$$

We conclude:

Theorem 13.3.1 ***Asymptotic Distribution of GMM Estimator***

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}}\right),$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{\Omega}\boldsymbol{W}\boldsymbol{Q}\right)\left(\boldsymbol{Q}'\boldsymbol{W}\boldsymbol{Q}\right)^{-1}.$$

In general, GMM estimators are asymptotically normal with "sandwich form" asymptotic variances.

The 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 as an exercise. This yields the *efficient GMM* estimator:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Omega}^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{\Omega}^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

Thus we have

Theorem 13.3.2 ***Asymptotic Distribution of Efficient GMM Estimator***

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\boldsymbol{0}, \left(\boldsymbol{Q}'\boldsymbol{\Omega}^{-1}\boldsymbol{Q}\right)^{-1}\right).$$

$\boldsymbol{W}_0 = \boldsymbol{\Omega}^{-1}$ is not known in practice, but it can be estimated consistently. For any $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{W}_0$, we still call $\widehat{\boldsymbol{\beta}}$ the efficient GMM estimator, as it has the same asymptotic distribution.

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 $\boldsymbol{W}_n$. However, it turns out that the GMM estimator is semiparametrically efficient, as shown by Gary Chamberlain (1987).

If it is known that $\mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta})\right) = \boldsymbol{0}$, 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 $\boldsymbol{G} = \mathbb{E}\frac{\partial}{\partial\boldsymbol{\beta}'}\boldsymbol{g}_i(\boldsymbol{\beta})$. Since the GMM estimator has this asymptotic variance, it is semiparametrically efficient.

This result shows that in the linear model, no estimator has greater asymptotic efficiency than the efficient linear GMM estimator. No estimator can do better (in this first-order asymptotic sense), without imposing additional assumptions.

13.4 Estimation of the Efficient Weight Matrix

Given any weight matrix $\boldsymbol{W}_n > 0$, the GMM estimator $\widehat{\boldsymbol{\beta}}$ is consistent yet inefficient. For example, we can set $\boldsymbol{W}_n = \boldsymbol{I}_\ell$. In the linear model, a better choice is $\boldsymbol{W}_n = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$. Given any such first-step estimator, we can define the residuals $\hat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ and moment equations $\hat{\boldsymbol{g}}_i = \boldsymbol{z}_i\hat{e}_i = \boldsymbol{g}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \widehat{\boldsymbol{\beta}})$. Construct

$$\overline{\boldsymbol{g}}_n = \overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{g}}_i,$$

$$\hat{\boldsymbol{g}}_i^* = \hat{\boldsymbol{g}}_i - \overline{\boldsymbol{g}}_n,$$

and define

$$\boldsymbol{W}_n = \left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{g}}_i^*\hat{\boldsymbol{g}}_i^{*\prime}\right)^{-1} = \left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{g}}_i\hat{\boldsymbol{g}}_i' - \overline{\boldsymbol{g}}_n\overline{\boldsymbol{g}}_n'\right)^{-1}. \tag{13.4}$$

Then $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{\Omega}^{-1} = \boldsymbol{W}_0$, and GMM using $\boldsymbol{W}_n$ as the weight matrix is asymptotically efficient.

A common alternative choice is to set

$$\boldsymbol{W}_n = \left(\frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{g}}_i\hat{\boldsymbol{g}}_i'\right)^{-1}$$

which uses the uncentered moment conditions. Since $\mathbb{E}\boldsymbol{g}_i = \boldsymbol{0}$, these two estimators are asymptotically equivalent under the hypothesis of correct specification. However, Alastair Hall (2000) has shown that the uncentered estimator is a poor choice. When constructing hypothesis tests, under the alternative hypothesis the moment conditions are violated, i.e. $\mathbb{E}\boldsymbol{g}_i \neq \boldsymbol{0}$, so the uncentered estimator will contain an undesirable bias term and the power of the test will be adversely affected. A simple solution is to use the centered moment conditions to construct the weight matrix, as in (13.4) above.

Here is a simple way to compute the efficient GMM estimator for the linear model. First, set $\boldsymbol{W}_n = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$, estimate $\widehat{\boldsymbol{\beta}}$ using this weight matrix, and construct the residual $\hat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$. Then set $\hat{\boldsymbol{g}}_i = \boldsymbol{z}_i\hat{e}_i$, and let $\hat{\boldsymbol{g}}$ be the associated $n \times \ell$ matrix. Then the efficient GMM estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{Z}\left(\hat{\boldsymbol{g}}'\hat{\boldsymbol{g}} - n\overline{\boldsymbol{g}}_n\overline{\boldsymbol{g}}_n'\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\hat{\boldsymbol{g}}'\hat{\boldsymbol{g}} - n\overline{\boldsymbol{g}}_n\overline{\boldsymbol{g}}_n'\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

In most cases, when we say "GMM", we actually mean "efficient GMM". There is little point in using an inefficient GMM estimator when the efficient estimator is easy to compute.

An estimator of the asymptotic variance of $\hat{\boldsymbol{\beta}}$ can be seen from the above formula. Set

$$\widehat{\boldsymbol{V}} = n\left(\boldsymbol{X}'\boldsymbol{Z}\left(\hat{\boldsymbol{g}}'\hat{\boldsymbol{g}} - n\overline{\boldsymbol{g}}_n\overline{\boldsymbol{g}}_n'\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}.$$

Asymptotic standard errors are given by the square roots of the diagonal elements of $\frac{1}{n}\widehat{\boldsymbol{V}}$.

There is an important alternative to the two-step GMM estimator just described. Instead, we can let the weight matrix be considered as a function of $\boldsymbol{\beta}$. The criterion function is then

$$J(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})'\left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{g}_i^*(\boldsymbol{\beta})\boldsymbol{g}_i^*(\boldsymbol{\beta})'\right)^{-1}\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}).$$

where

$$\boldsymbol{g}_i^*(\boldsymbol{\beta}) = \boldsymbol{g}_i(\boldsymbol{\beta}) - \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$$

The $\widehat{\boldsymbol{\beta}}$ which minimizes this function is called the **continuously-updated GMM estimator**, and was introduced by L. Hansen, Heaton and Yaron (1996).

The estimator appears to have some better properties than traditional GMM, but can be numerically tricky to obtain in some cases. This is a current area of research in econometrics.

13.5 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 $l \geq k = \dim(\boldsymbol{\beta})$. The GMM estimator minimizes

$$J(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \boldsymbol{W}_n \, \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$$

where

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\beta})$$

and

$$\boldsymbol{W}_n = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{g}}_i \hat{\boldsymbol{g}}_i' - \overline{\boldsymbol{g}}_n \overline{\boldsymbol{g}}_n'\right)^{-1},$$

with $\hat{\boldsymbol{g}}_i = \boldsymbol{g}_i(\widetilde{\boldsymbol{\beta}})$ constructed using a preliminary consistent estimator $\widetilde{\boldsymbol{\beta}}$, perhaps obtained by first setting $\boldsymbol{W}_n = \boldsymbol{I}$. Since the GMM estimator depends upon the first-stage estimator, often the weight matrix $\boldsymbol{W}_n$ is updated, and then $\widehat{\boldsymbol{\beta}}$ recomputed. This estimator can be iterated if needed.

Theorem 13.5.1 ***Distribution of Nonlinear GMM Estimator***
Under general regularity conditions,

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\right),$$

where

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{g}_i \boldsymbol{g}_i'\right)$$

and

$$\boldsymbol{G} = \mathbb{E}\frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\boldsymbol{\beta}).$$

The variance of $\widehat{\boldsymbol{\beta}}$ *may be estimated by*

$$\widehat{\boldsymbol{V}}_{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{G}}'\hat{\boldsymbol{\Omega}}^{-1}\hat{\boldsymbol{G}}\right)^{-1}$$

where

$$\hat{\boldsymbol{\Omega}} = n^{-1}\sum_i \hat{\boldsymbol{g}}_i^* \hat{\boldsymbol{g}}_i^{*\prime}$$

and

$$\hat{\boldsymbol{G}} = n^{-1}\sum_i \frac{\partial}{\partial \boldsymbol{\beta}'}\boldsymbol{g}_i(\hat{\boldsymbol{\beta}}).$$

The general theory of GMM estimation and testing was exposited by L. Hansen (1982).

13.6 Over-Identification Test

Overidentified models ($\ell > k$) are special in the sense that there may not be a parameter value $\boldsymbol{\beta}$ such that the moment condition

$$\mathbb{E}\boldsymbol{g}(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i, \boldsymbol{\beta}) = \mathbf{0}$$

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 $\mathbb{E}(\boldsymbol{x}_{1i} e_i) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{x}_{\boldsymbol{2i}} e_i) = \mathbf{0}$. 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 $\mathbb{E}(\boldsymbol{x}_{1i}(y_i - \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1)) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{x}_{2i}(y_i - \boldsymbol{x}_{1i}'\boldsymbol{\beta}_1)) = \mathbf{0}$ hold simultaneously. In this sense an exclusion restriction can be seen as an overidentifying restriction.

Note that $\overline{\boldsymbol{g}}_n \xrightarrow{p} \mathbb{E}\boldsymbol{g}_i$, and thus $\overline{\boldsymbol{g}}_n$ can be used to assess whether or not the hypothesis that $\mathbb{E}\boldsymbol{g}_i = \mathbf{0}$ is true or not. The criterion function at the parameter estimates is

$$\begin{aligned} J_n &= n\, \overline{\boldsymbol{g}}_n' \boldsymbol{W}_n \overline{\boldsymbol{g}}_n \\ &= n^2 \overline{\boldsymbol{g}}_n' \left(\hat{\boldsymbol{g}}' \hat{\boldsymbol{g}} - n \overline{\boldsymbol{g}}_n \overline{\boldsymbol{g}}_n' \right)^{-1} \overline{\boldsymbol{g}}_n. \end{aligned}$$

is a quadratic form in $\overline{\boldsymbol{g}}_n$, and is thus a natural test statistic for $\mathbb{H}_0 : \mathbb{E}\boldsymbol{g}_i = \mathbf{0}$.

Theorem 13.6.1 *(Sargan-Hansen). Under the hypothesis of correct specification, and if the weight matrix is asymptotically efficient,*

$$J_n = J_n(\widehat{\boldsymbol{\beta}}) \xrightarrow{d} \chi^2_{\ell-k}.$$

The proof of the theorem is left as an exercise. This result was established by Sargan (1958) for a specialized case, and by L. Hansen (1982) for the general case.

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.

13.7 Hypothesis Testing: The Distance Statistic

We described before how to construct estimates of the asymptotic covariance matrix of the GMM estimates. These may be used to construct Wald tests of statistical hypotheses.

If the hypothesis is non-linear, a better approach is to directly use the GMM criterion function. 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).

For a given weight matrix $\boldsymbol{W}_n$, the GMM criterion function is

$$J_n(\boldsymbol{\beta}) = n \cdot \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})' \boldsymbol{W}_n \, \overline{\boldsymbol{g}}_n(\boldsymbol{\beta})$$

For $\boldsymbol{h} : \mathbb{R}^k \to \mathbb{R}^r$, the hypothesis is

$$\mathbb{H}_0 : \boldsymbol{h}(\boldsymbol{\beta}) = \mathbf{0}.$$

The estimates under $\mathbb{H}_1$ are

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, J_n(\boldsymbol{\beta})$$

and those under $\mathbb{H}_0$ are

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{h}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmin}} J(\boldsymbol{\beta}).$$

The two minimizing criterion functions are $J_n(\widehat{\boldsymbol{\beta}})$ and $J_n(\widetilde{\boldsymbol{\beta}})$. The GMM distance statistic is the difference

$$D_n = J_n(\widetilde{\boldsymbol{\beta}}) - J_n(\widehat{\boldsymbol{\beta}}).$$

Proposition 13.7.1 *If the same weight matrix $\boldsymbol{W}_n$ is used for both null and alternative,*

1. *$D \geq 0$*
2. *$D \xrightarrow{d} \chi^2_r$*
3. *If $\boldsymbol{h}$ is linear in $\boldsymbol{\beta}$, then D equals the Wald statistic.*

If $\boldsymbol{h}$ is non-linear, the Wald statistic can work quite poorly. In contrast, current evidence suggests that the D_n statistic appears to have quite good sampling properties, and is the preferred test statistic.

Newey and West (1987) suggested to use the same weight matrix $\boldsymbol{W}_n$ for both null and alternative, as this ensures that $D_n \geq 0$. This reasoning is not compelling, however, and some current research suggests that this restriction is not necessary for good performance of the test.

This test shares the useful feature of LR tests in that it is a natural by-product of the computation of alternative models.

13.8 Conditional Moment Restrictions

In many contexts, the model implies more than an unconditional moment restriction of the form $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0}$. 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 restriction discussed above.

As discussed later in Chapter 15, the linear model $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ with instruments $\boldsymbol{z}_i$ falls into this class under the assumption $\mathbb{E}\left(e_i \mid \boldsymbol{z}_i\right) = 0$. 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 $\mathbb{E}\left(y \mid \boldsymbol{x}\right) = \boldsymbol{x}'\boldsymbol{\beta}$ and variance $\operatorname{var}\left(y \mid \boldsymbol{x}\right) = f\left(\boldsymbol{x}\right)'\boldsymbol{\gamma}$, then

$$\boldsymbol{e}_i\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \left\{ \begin{array}{c} y_i - \boldsymbol{x}_i'\boldsymbol{\beta} \\ \\ \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2 - f\left(\boldsymbol{x}_i\right)'\boldsymbol{\gamma} \end{array} \right. .$$

Here $s = 2$.

Given a conditional moment restriction, an unconditional moment restriction can always be constructed. That is for any $\ell \times 1$ function $\boldsymbol{\phi}\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right)$, we can set $\boldsymbol{g}_i(\boldsymbol{\beta}) = \boldsymbol{\phi}\left(\boldsymbol{x}_i, \boldsymbol{\beta}\right) e_i(\boldsymbol{\beta})$ which satisfies $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0}$ and hence defines a GMM estimator. 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 $x_i \in \mathbb{R}$ is a valid instrument satisfying $\mathbb{E}(e_i \mid x_i) = 0$, then x_i, x_i^2, x_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 GLS, as we discussed earlier in the course.

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.

13.9 Bootstrap GMM Inference

Let $\widehat{\boldsymbol{\beta}}$ be the 2SLS or GMM estimator of $\boldsymbol{\beta}$. Using the EDF of $(y_i, \boldsymbol{z}_i, \boldsymbol{x}_i)$, we can apply the bootstrap methods discussed in Chapter 10 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 $\overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) \neq 0$. Thus according to random variables $(y_i^*, \boldsymbol{z}_i^*, \boldsymbol{x}_i^*)$ drawn from the EDF F_n,

$$\mathbb{E}\left(\boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right)\right) = \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}}) \neq \mathbf{0}.$$

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

$$J_n^*(\boldsymbol{\beta}) = n \cdot \left(\overline{\boldsymbol{g}}_n^*(\boldsymbol{\beta}) - \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})\right)' \boldsymbol{W}_n^* \left(\overline{\boldsymbol{g}}_n^*(\boldsymbol{\beta}) - \overline{\boldsymbol{g}}_n(\hat{\boldsymbol{\beta}})\right)$$

where $\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})$ is from the in-sample data, not from the bootstrap data.

Let $\widehat{\boldsymbol{\beta}}^*$ minimize $J_n^*(\boldsymbol{\beta})$, and define all statistics and tests accordingly. In the linear model, this implies that the bootstrap estimator is

$$\widehat{\boldsymbol{\beta}}_n^* = \left(\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\boldsymbol{W}_n^*\boldsymbol{Z}^{*\prime}\boldsymbol{X}^*\right)^{-1}\left(\boldsymbol{X}^{*\prime}\boldsymbol{Z}^*\boldsymbol{W}_n^*\left(\boldsymbol{Z}^{*\prime}\boldsymbol{y}^* - \boldsymbol{Z}'\hat{\boldsymbol{e}}\right)\right).$$

where $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ are the in-sample residuals. The bootstrap J statistic is $J_n^*(\widehat{\boldsymbol{\beta}}^*)$.

Brown and Newey (2002) have an alternative solution. They note that we can sample from the observations with the empirical likelihood probabilities $\hat{p}_i$ described in Chapter 14. Since $\sum_{i=1}^n \hat{p}_i \boldsymbol{g}_i\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{0}$, this sampling scheme preserves the moment conditions of the model, so no recentering or adjustments is needed. Brown and Newey argue that this bootstrap procedure will be more efficient than the Hall-Horowitz GMM bootstrap.

Exercises

Exercise 13.1 Take the model

$$\begin{aligned} y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\ \mathbb{E}(\boldsymbol{x}_i e_i) &= \mathbf{0} \\ e_i^2 &= \boldsymbol{z}_i'\boldsymbol{\gamma} + \eta_i \\ \mathbb{E}(\boldsymbol{z}_i \eta_i) &= \mathbf{0}. \end{aligned}$$

Find the method of moments estimators $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}\right)$ for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.

Exercise 13.2 Take the single equation

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \\ \mathbb{E}(\boldsymbol{e} \mid \boldsymbol{Z}) &= \mathbf{0} \end{aligned}$$

Assume $\mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right) = \sigma^2$. Show that if $\hat{\boldsymbol{\beta}}$ is estimated by GMM with weight matrix $\boldsymbol{W}_n = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}$, 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 $\boldsymbol{Q} = \mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i')$ and $\boldsymbol{M} = \mathbb{E}(\boldsymbol{z}_i\boldsymbol{z}_i')$.

Exercise 13.3 Take the model $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ with $\mathbb{E}(\boldsymbol{z}_i e_i) = \mathbf{0}$. Let $\hat{e}_i = y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$ where $\widehat{\boldsymbol{\beta}}$ is consistent for $\boldsymbol{\beta}$ (e.g. a GMM estimator with arbitrary weight matrix). Define the estimate of the optimal GMM weight matrix

$$\boldsymbol{W}_n = \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i\boldsymbol{z}_i'\hat{e}_i^2\right)^{-1}.$$

Show that $\boldsymbol{W}_n \xrightarrow{p} \boldsymbol{\Omega}^{-1}$ where $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)$.

Exercise 13.4 In the linear model estimated by GMM with general weight matrix $\boldsymbol{W}$, the asymptotic variance of $\hat{\boldsymbol{\beta}}_{GMM}$ 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 $\boldsymbol{B}'\boldsymbol{\Omega}(\boldsymbol{A} - \boldsymbol{B}) = \mathbf{0}$.

(d) Use the expressions $\boldsymbol{V} = \boldsymbol{A}'\boldsymbol{\Omega}\boldsymbol{A}$, $\boldsymbol{A} = \boldsymbol{B} + (\boldsymbol{A} - \boldsymbol{B})$, and $\boldsymbol{B}'\boldsymbol{\Omega}(\boldsymbol{A} - \boldsymbol{B}) = \mathbf{0}$ to show that $\boldsymbol{V} \geq \boldsymbol{V}_0$.

Exercise 13.5 The equation of interest is

$$\begin{aligned} y_i &= \boldsymbol{m}(\boldsymbol{x}_i, \boldsymbol{\beta}) + e_i \\ \mathbb{E}(\boldsymbol{z}_i e_i) &= \mathbf{0}. \end{aligned}$$

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 13.6 In the linear model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}$ with $\mathbb{E}(\boldsymbol{x}_i e_i) = \boldsymbol{0}$, a Generalized Method of Moments (GMM) criterion function for $\boldsymbol{\beta}$ is defined as

$$J_n(\boldsymbol{\beta}) = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X\beta})' \boldsymbol{X} \widehat{\boldsymbol{\Omega}}^{-1} \boldsymbol{X}' (\boldsymbol{y} - \boldsymbol{X\beta}) \tag{13.5}$$

where $\widehat{\boldsymbol{\Omega}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i' \hat{e}_i^2$, $\hat{e}_i = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ 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 $\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{0}$, is defined as

$$\widetilde{\boldsymbol{\beta}} = \underset{\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{0}}{\operatorname{argmin}} J_n(\boldsymbol{\beta}).$$

The GMM test statistic (the distance statistic) of the hypothesis $\boldsymbol{h}(\boldsymbol{\beta}) = \boldsymbol{0}$ is

$$D = J_n(\widetilde{\boldsymbol{\beta}}) = \min_{\boldsymbol{h}(\boldsymbol{\beta})=\boldsymbol{0}} J_n(\boldsymbol{\beta}). \tag{13.6}$$

(a) Show that you can rewrite $J_n(\boldsymbol{\beta})$ in (13.5) as

$$J_n(\boldsymbol{\beta}) = n\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)' \hat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)$$

thus $\widetilde{\boldsymbol{\beta}}$ is the same as the minimum distance estimator.

(b) Show that under linear hypotheses the distance statistic D in (13.6) equals the Wald statistic.

Exercise 13.7 Take the linear model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$$
$$\mathbb{E}(\boldsymbol{z}_i e_i) = \boldsymbol{0}.$$

and consider the GMM estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. Let

$$J_n = n\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})'\widehat{\boldsymbol{\Omega}}^{-1}\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})$$

denote the test of overidentifying restrictions. Show that $J_n \xrightarrow{d} \chi^2_{\ell-k}$ as $n \to \infty$ by demonstrating each of the following:

(a) Since $\boldsymbol{\Omega} > 0$, we can write $\boldsymbol{\Omega}^{-1} = \boldsymbol{CC}'$ and $\boldsymbol{\Omega} = \boldsymbol{C}'^{-1}\boldsymbol{C}^{-1}$

(b) $J_n = n\left(\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})\right)'\left(\boldsymbol{C}'\widehat{\boldsymbol{\Omega}}\boldsymbol{C}\right)^{-1}\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}})$

(c) $\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\widehat{\boldsymbol{\beta}}) = \boldsymbol{D}_n\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0)$ where

$$\boldsymbol{D}_n = \boldsymbol{I}_\ell - \boldsymbol{C}'\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\left(\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\widehat{\boldsymbol{\Omega}}^{-1}\left(\frac{1}{n}\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{Z}\right)\widehat{\boldsymbol{\Omega}}^{-1}\boldsymbol{C}'^{-1}$$

$$\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) = \frac{1}{n}\boldsymbol{Z}'\boldsymbol{e}.$$

(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) $n^{1/2}\boldsymbol{C}'\overline{\boldsymbol{g}}_n(\boldsymbol{\beta}_0) \xrightarrow{d} \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_\ell)$

(f) $J_n \xrightarrow{d} \boldsymbol{u}'\left(\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\right)\boldsymbol{u}$

(g) $\boldsymbol{u}'\left(\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'\right)\boldsymbol{u} \sim \chi^2_{\ell-k}$.

Hint: $\boldsymbol{I}_\ell - \boldsymbol{R}(\boldsymbol{R}'\boldsymbol{R})^{-1}\boldsymbol{R}'$ is a projection matrix.

Chapter 14

Empirical Likelihood

14.1 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

$$\sum_{i=1}^{n} p_i = 1. \tag{14.1}$$

Since each observation is observed once in the sample, the log-likelihood function for this multinomial distribution is

$$\log L\left(p_1,...,p_n\right) = \sum_{i=1}^{n} \log(p_i). \tag{14.2}$$

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 (14.1). 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 (14.1) 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

$$\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0) = \mathbf{0}$$

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

$$\sum_{i=1}^{n} p_i \boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0} \tag{14.3}$$

The **empirical likelihood estimator** is the value of $\boldsymbol{\beta}$ which maximizes the multinomial log-likelihood (14.2) subject to the restrictions (14.1) and (14.3).

The Lagrangian for this maximization problem is

$$\mathcal{L}\left(\boldsymbol{\beta}, p_1, ..., p_n, \boldsymbol{\lambda}, \mu\right) = \sum_{i=1}^{n} \log(p_i) - \mu \left(\sum_{i=1}^{n} p_i - 1 \right) - n\boldsymbol{\lambda}' \sum_{i=1}^{n} p_i \boldsymbol{g}_i\left(\boldsymbol{\beta}\right)$$

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

$$\begin{aligned} \frac{1}{p_i} &= \mu + n\boldsymbol{\lambda}'\boldsymbol{g}_i\left(\boldsymbol{\beta}\right) \\ \sum_{i=1}^{n} p_i &= 1 \\ \sum_{i=1}^{n} p_i \boldsymbol{g}_i\left(\boldsymbol{\beta}\right) &= \mathbf{0}. \end{aligned}$$

Multiplying the first equation by p_i, summing over i, and using the second and third equations, we find $\mu = n$ and

$$p_i = \frac{1}{n\left(1 + \boldsymbol{\lambda}'\boldsymbol{g}_i\left(\boldsymbol{\beta}\right)\right)}.$$

Substituting into $\mathcal{L}$ we find

$$R\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right) = -n\log\left(n\right) - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}'\boldsymbol{g}_i\left(\boldsymbol{\beta}\right)\right). \tag{14.4}$$

For given $\boldsymbol{\beta}$, the Lagrange multiplier $\boldsymbol{\lambda}(\boldsymbol{\beta})$ minimizes $R\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right)$:

$$\boldsymbol{\lambda}(\boldsymbol{\beta}) = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}}\, R(\boldsymbol{\beta}, \boldsymbol{\lambda}). \tag{14.5}$$

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 $\boldsymbol{\beta}$.

$$\begin{aligned} R(\boldsymbol{\beta}) &= R(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})) \\ &= -n\log\left(n\right) - \sum_{i=1}^{n} \log\left(1 + \boldsymbol{\lambda}(\boldsymbol{\beta})'\boldsymbol{g}_i\left(\boldsymbol{\beta}\right)\right) \end{aligned}$$

The EL estimate $\hat{\boldsymbol{\beta}}$ is the value which maximizes $R(\boldsymbol{\beta})$, or equivalently minimizes its negative

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left[-R(\boldsymbol{\beta})\right] \tag{14.6}$$

Numerical methods are required for calculation of $\hat{\boldsymbol{\beta}}$ (see Section 14.5).

As a by-product of estimation, we also obtain the Lagrange multiplier $\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}(\hat{\boldsymbol{\beta}})$, probabilities

$$\hat{p}_i = \frac{1}{n\left(1 + \hat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\right)}.$$

and maximized empirical likelihood

$$R(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \log\left(\hat{p}_i\right). \tag{14.7}$$

14.2 Asymptotic Distribution of EL Estimator

Let $\boldsymbol{\beta}_0$ denote the true value of $\boldsymbol{\beta}$ and define

$$\boldsymbol{G}_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}'} \boldsymbol{g}_i(\boldsymbol{\beta}) \tag{14.8}$$

$$\boldsymbol{G} = \mathbb{E}\boldsymbol{G}_i(\boldsymbol{\beta}_0)$$

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{g}_i(\boldsymbol{\beta}_0)\boldsymbol{g}_i(\boldsymbol{\beta}_0)'\right)$$

and

$$\boldsymbol{V} = \left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1} \tag{14.9}$$

$$\boldsymbol{V}_{\boldsymbol{\lambda}} = \boldsymbol{\Omega} - \boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}' \tag{14.10}$$

For example, in the linear model, $\boldsymbol{G}_i(\boldsymbol{\beta}) = -\boldsymbol{z}_i\boldsymbol{x}_i'$, $\boldsymbol{G} = -\mathbb{E}(\boldsymbol{z}_i\boldsymbol{x}_i')$, and $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)$.

Theorem 14.2.1 *Under regularity conditions,*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\beta}})$$

$$\sqrt{n}\hat{\boldsymbol{\lambda}} \xrightarrow{d} \boldsymbol{\Omega}^{-1}\mathrm{N}(\boldsymbol{0}, \boldsymbol{V}_{\boldsymbol{\lambda}})$$

where $\boldsymbol{V}$ *and* $\boldsymbol{V}_{\boldsymbol{\lambda}}$ *are defined in (14.9) and (14.10), and* $\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)$ *and* $\sqrt{n}\hat{\boldsymbol{\lambda}}$ *are asymptotically independent.*

The theorem shows that the asymptotic variance $\boldsymbol{V}_{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}$ 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 14.2.1. $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\lambda}})$ jointly solve

$$\boldsymbol{0} = \frac{\partial}{\partial \boldsymbol{\lambda}} R(\boldsymbol{\beta}, \boldsymbol{\lambda}) = -\sum_{i=1}^{n} \frac{\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)}{\left(1 + \hat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\right)} \tag{14.11}$$

$$\boldsymbol{0} = \frac{\partial}{\partial \boldsymbol{\beta}} R(\boldsymbol{\beta}, \boldsymbol{\lambda}) = -\sum_{i=1}^{n} \frac{\boldsymbol{G}_i\left(\hat{\boldsymbol{\beta}}\right)'\boldsymbol{\lambda}}{1 + \hat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)}. \tag{14.12}$$

Let $\boldsymbol{G}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{G}_i(\boldsymbol{\beta}_0)$, $\overline{\boldsymbol{g}}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta}_0)$ and $\boldsymbol{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta}_0)\boldsymbol{g}_i(\boldsymbol{\beta}_0)'$.

Expanding (14.12) around $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 = \boldsymbol{0}$ yields

$$\boldsymbol{0} \simeq \boldsymbol{G}_n'\left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}_0\right). \tag{14.13}$$

Expanding (14.11) around $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0 = \boldsymbol{0}$ yields

$$\boldsymbol{0} \simeq -\overline{\boldsymbol{g}}_n - \boldsymbol{G}_n\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) + \boldsymbol{\Omega}_n\hat{\boldsymbol{\lambda}} \tag{14.14}$$

Premultiplying by $\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}$ and using (14.13) yields

$$\begin{aligned}\mathbf{0} &\simeq -\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\overline{\boldsymbol{g}}_n - \boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)+\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{\Omega}_n\hat{\boldsymbol{\lambda}}\\ &= -\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\overline{\boldsymbol{g}}_n - \boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\end{aligned}$$

Solving for $\hat{\boldsymbol{\beta}}$ and using the WLLN and CLT yields

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right) &\simeq -\left(\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\sqrt{n}\overline{\boldsymbol{g}}_n \\ &\xrightarrow{d}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right)\\ &= \mathrm{N}\left(\mathbf{0},\,\boldsymbol{V}_{\boldsymbol{\beta}}\right)\end{aligned} \tag{14.15}$$

Solving (14.14) for $\hat{\boldsymbol{\lambda}}$ and using (14.15) yields

$$\begin{aligned}\sqrt{n}\hat{\boldsymbol{\lambda}} &\simeq \boldsymbol{\Omega}_n^{-1}\left(\boldsymbol{I}-\boldsymbol{G}_n\left(\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\right)\sqrt{n}\overline{\boldsymbol{g}}_n \\ &\xrightarrow{d}\boldsymbol{\Omega}^{-1}\left(\boldsymbol{I}-\boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\right)\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right)\\ &= \boldsymbol{\Omega}^{-1}\mathrm{N}\left(\mathbf{0},\,\boldsymbol{V}_{\boldsymbol{\lambda}}\right)\end{aligned} \tag{14.16}$$

Furthermore, since

$$\boldsymbol{G}'\left(\boldsymbol{I}-\boldsymbol{\Omega}^{-1}\boldsymbol{G}\left(\boldsymbol{G}'\boldsymbol{\Omega}^{-1}\boldsymbol{G}\right)^{-1}\boldsymbol{G}'\right)=\mathbf{0}$$

$\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)$ and $\sqrt{n}\hat{\boldsymbol{\lambda}}$ are asymptotically uncorrelated and hence independent.

14.3 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 (14.7) is

$$LR_n=\sum_{i=1}^n 2\log\left(1+\hat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\right). \tag{14.17}$$

Theorem 14.3.1 *If* $\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}_0)=\mathbf{0}$ *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 14.3.1. First, by a Taylor expansion, (14.15), and (14.16),

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right) &\simeq \sqrt{n}\left(\overline{\boldsymbol{g}}_n+\boldsymbol{G}_n\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\right)\right)\\ &\simeq\left(\boldsymbol{I}-\boldsymbol{G}_n\left(\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\boldsymbol{G}_n\right)^{-1}\boldsymbol{G}_n'\boldsymbol{\Omega}_n^{-1}\right)\sqrt{n}\overline{\boldsymbol{g}}_n\\ &\simeq\boldsymbol{\Omega}_n\sqrt{n}\hat{\boldsymbol{\lambda}}.\end{aligned}$$

Second, since $\log(1+u) \simeq u - u^2/2$ for u small,

$$\begin{aligned}
LR_n &= \sum_{i=1}^{n} 2\log\left(1+\hat{\boldsymbol{\lambda}}'\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\right) \\
&\simeq 2\hat{\boldsymbol{\lambda}}'\sum_{i=1}^{n}\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right) - \hat{\boldsymbol{\lambda}}'\sum_{i=1}^{n}\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{g}_i\left(\hat{\boldsymbol{\beta}}\right)'\hat{\boldsymbol{\lambda}} \\
&\simeq n\hat{\boldsymbol{\lambda}}'\boldsymbol{\Omega}_n\hat{\boldsymbol{\lambda}} \\
&\xrightarrow{d} \mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\lambda}}\right)'\boldsymbol{\Omega}^{-1}\mathrm{N}\left(\mathbf{0},\boldsymbol{V}_{\boldsymbol{\lambda}}\right) \\
&= \chi^2_{\ell-k}
\end{aligned}$$

where the proof of the final equality is left as an exercise.

14.4 Testing

Let the maintained model be

$$\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}) = \mathbf{0} \tag{14.18}$$

where $\boldsymbol{g}$ is $\ell \times 1$ and $\boldsymbol{\beta}$ is $k \times 1$. By "maintained" we mean that the overidentfying restrictions contained in (14.18) 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

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \underset{\boldsymbol{h}(\boldsymbol{\beta})=\mathbf{0}}{\operatorname{argmax}} R(\boldsymbol{\beta}) \\
R(\tilde{\boldsymbol{\beta}}) &= \max_{\boldsymbol{h}(\boldsymbol{\beta})=\mathbf{0}} R(\boldsymbol{\beta}).
\end{aligned}$$

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 $\tilde{\boldsymbol{\beta}}$ relative to $\hat{\boldsymbol{\beta}}$. To test the hypothesis $\boldsymbol{h}(\boldsymbol{\beta})$ while maintaining (14.18), the simple overidentifying restrictions test (14.17) is not appropriate. Instead we use the difference in log-likelihoods:

$$LR_n = 2\left(R(\hat{\boldsymbol{\beta}}) - R(\tilde{\boldsymbol{\beta}})\right).$$

This test statistic is a natural analog of the GMM distance statistic.

Theorem 14.4.1 *Under (14.18) 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.

14.5 Numerical Computation

Gauss code which implements the methods discussed below can be found at

```
http://www.ssc.wisc.edu/~bhansen/progs/elike.prc
```

Derivatives
The numerical calculations depend on derivatives of the dual likelihood function (14.4). 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 (14.4) 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

$$\boldsymbol{R}_{\boldsymbol{\lambda\lambda}} = \frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^n \boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})\,\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})'$$
$$\boldsymbol{R}_{\boldsymbol{\lambda\beta}} = \frac{\partial^2}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\beta}'}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^n \left(\boldsymbol{g}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})\,\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})' - \frac{\boldsymbol{G}_i(\boldsymbol{\beta})}{1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})}\right)$$
$$\boldsymbol{R}_{\boldsymbol{\beta\beta}} = \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}R(\boldsymbol{\beta},\boldsymbol{\lambda}) = \sum_{i=1}^n \left(\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})\,\boldsymbol{G}_i^*(\boldsymbol{\beta},\boldsymbol{\lambda})' - \frac{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'}\left(\boldsymbol{g}_i(\boldsymbol{\beta})'\boldsymbol{\lambda}\right)}{1+\boldsymbol{\lambda}'\boldsymbol{g}_i(\boldsymbol{\beta})}\right)$$

Inner Loop
The so-called "inner loop" solves (14.5) for given $\boldsymbol{\beta}$. The modified Newton method takes a quadratic approximation to $R_n(\boldsymbol{\beta},\boldsymbol{\lambda})$ yielding the iteration rule

$$\boldsymbol{\lambda}_{j+1} = \boldsymbol{\lambda}_j - \delta\left(\boldsymbol{R}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j)\right)^{-1}\boldsymbol{R}_{\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j). \tag{14.19}$$

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 (14.19) 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

$$\boldsymbol{\lambda}_p = \boldsymbol{\lambda}_j - \delta_p\left(\boldsymbol{R}_{\boldsymbol{\lambda\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j)\right)^{-1}\boldsymbol{R}_{\boldsymbol{\lambda}}(\boldsymbol{\beta},\boldsymbol{\lambda}_j))$$
$$R_p = R(\boldsymbol{\beta},\boldsymbol{\lambda}_p)$$

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 (14.19).

A complication is that $\boldsymbol{\lambda}$ must be constrained so that $0 \leq p_i \leq 1$ which holds if

$$n\left(1+\boldsymbol{\lambda}'\boldsymbol{g}_i\left(\boldsymbol{\beta}\right)\right) \geq 1 \tag{14.20}$$

for all i. If (14.20) fails, the stepsize δ needs to be decreased.

Outer Loop

The outer loop is the minimization (14.6). This can be done by the modified Newton method described in the previous section. The gradient for (14.6) 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 $\boldsymbol{R}_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}\right) = 0$ 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 $\boldsymbol{R}_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})\right) = 0$.

The Hessian for (14.6) is

$$\begin{aligned}
\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} R(\boldsymbol{\beta}) \\
&= -\frac{\partial}{\partial \boldsymbol{\beta}'} \left[\boldsymbol{R}_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})\right) + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})\right)\right] \\
&= -\left(\boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}\left(\boldsymbol{\beta}, \boldsymbol{\lambda}(\boldsymbol{\beta})\right) + \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}}' \boldsymbol{\lambda}_{\boldsymbol{\beta}} + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}} + \boldsymbol{\lambda}_{\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}} \boldsymbol{\lambda}_{\boldsymbol{\beta}}\right) \\
&= \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}}' \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\lambda}}^{-1} \boldsymbol{R}_{\boldsymbol{\lambda}\boldsymbol{\beta}} - \boldsymbol{R}_{\boldsymbol{\beta}\boldsymbol{\beta}}.
\end{aligned}$$

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 15

Endogeneity

We say that there is endogeneity in the linear model $y = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i$ if $\boldsymbol{\beta}$ is the parameter of interest and $\mathbb{E}(\boldsymbol{x}_i e_i) \neq 0$. This cannot happen if $\boldsymbol{\beta}$ is defined by linear projection, so requires a structural interpretation. The coefficient $\boldsymbol{\beta}$ must have meaning separately from the definition of a conditional mean or linear projection.

Example: Measurement error in the regressor. Suppose that $(y_i, \boldsymbol{x}_i^*)$ are joint random variables, $\mathbb{E}(y_i \mid \boldsymbol{x}_i^*) = \boldsymbol{x}_i^{*\prime}\boldsymbol{\beta}$ is linear, $\boldsymbol{\beta}$ is the parameter of interest, and $\boldsymbol{x}_i^*$ is not observed. Instead we observe $\boldsymbol{x}_i = \boldsymbol{x}_i^* + \boldsymbol{u}_i$ where $\boldsymbol{u}_i$ is an $k \times 1$ measurement error, independent of y_i and $\boldsymbol{x}_i^*$. Then

$$\begin{aligned} y_i &= \boldsymbol{x}_i^{*\prime}\boldsymbol{\beta} + e_i \\ &= (\boldsymbol{x}_i - \boldsymbol{u}_i)'\,\boldsymbol{\beta} + e_i \\ &= \boldsymbol{x}_i'\boldsymbol{\beta} + v_i \end{aligned}$$

where

$$v_i = e_i - \boldsymbol{u}_i'\boldsymbol{\beta}.$$

The problem is that

$$\mathbb{E}\left(\boldsymbol{x}_i v_i\right) = \mathbb{E}\left[\left(\boldsymbol{x}_i^* + \boldsymbol{u}_i\right)\left(e_i - \boldsymbol{u}_i'\boldsymbol{\beta}\right)\right] = -\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)\boldsymbol{\beta} \neq 0$$

if $\boldsymbol{\beta} \neq 0$ and $\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right) \neq 0$. It follows that if $\hat{\boldsymbol{\beta}}$ is the OLS estimator, then

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}^* = \boldsymbol{\beta} - \left(\mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right)\right)^{-1}\mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right)\boldsymbol{\beta} \neq \boldsymbol{\beta}.$$

This is called **measurement error 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 iid, $\mathbb{E}\boldsymbol{e}_i = \boldsymbol{0}$ 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

$$\begin{aligned}\left(\begin{array}{c} q_i \\ p_i \end{array}\right) &= \left[\begin{array}{cc} 1 & \beta_1 \\ 1 & -\beta_2 \end{array}\right]^{-1} \left(\begin{array}{c} e_{1i} \\ e_{2i} \end{array}\right) \\ &= \left[\begin{array}{cc} \beta_2 & \beta_1 \\ 1 & -1 \end{array}\right] \left(\begin{array}{c} e_{1i} \\ e_{2i} \end{array}\right) \left(\frac{1}{\beta_1+\beta_2}\right) \\ &= \left(\begin{array}{c} (\beta_2 e_{1i} + \beta_1 e_{2i})/(\beta_1+\beta_2) \\ (e_{1i} - e_{2i})/(\beta_1+\beta_2) \end{array}\right).\end{aligned}$$

The projection of q_i on p_i yields

$$\begin{aligned} q_i &= \beta^* p_i + \varepsilon_i \\ \mathbb{E}(p_i \varepsilon_i) &= 0 \end{aligned}$$

where

$$\beta^* = \frac{\mathbb{E}(p_i q_i)}{\mathbb{E}(p_i^2)} = \frac{\beta_2 - \beta_1}{2}$$

Hence if it is estimated by OLS, $\hat{\beta} \xrightarrow{p} \beta^*$, which does not equal either β_1 or β_2. This is called **simultaneous equations bias**.

15.1 Instrumental Variables

Let the equation of interest be

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + e_i \tag{15.1}$$

where $\boldsymbol{x}_i$ is $k \times 1$, and assume that $\mathbb{E}(\boldsymbol{x}_i e_i) \neq 0$ so there is **endogeneity**. We call (15.1) the structural equation. In matrix notation, this can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}. \tag{15.2}$$

Any solution to the problem of endogeneity requires additional information which we call **instruments**.

Definition 15.1.1 *The $\ell \times 1$ random vector $\boldsymbol{z}_i$ is an* ***instrumental variable*** *for (15.1) if $\mathbb{E}(\boldsymbol{z}_i e_i) = \boldsymbol{0}$.*

In a typical set-up, some regressors in $\boldsymbol{x}_i$ will be uncorrelated with e_i (for example, at least the intercept). Thus we make the partition

$$\boldsymbol{x}_i = \left(\begin{array}{c} \boldsymbol{x}_{1i} \\ \boldsymbol{x}_{2i} \end{array}\right) \begin{array}{c} k_1 \\ k_2 \end{array} \tag{15.3}$$

where $\mathbb{E}(\boldsymbol{x}_{1i} e_i) = \boldsymbol{0}$ yet $\mathbb{E}(\boldsymbol{x}_{2i} e_i) \neq \boldsymbol{0}$. We call $\boldsymbol{x}_{1i}$ exogenous and $\boldsymbol{x}_{2i}$ endogenous. By the above definition, $\boldsymbol{x}_{1i}$ is an instrumental variable for (15.1), so should be included in $\boldsymbol{z}_i$. So we have the partition

$$\boldsymbol{z}_i = \left(\begin{array}{c} \boldsymbol{x}_{1i} \\ \boldsymbol{z}_{2i} \end{array}\right) \begin{array}{c} k_1 \\ \ell_2 \end{array} \tag{15.4}$$

where $\boldsymbol{x}_{1i} = \boldsymbol{z}_{1i}$ 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.

The model is **just-identified** if $\ell = k$ (i.e., if $\ell_2 = k_2$) and **over-identified** if $\ell > k$ (i.e., if $\ell_2 > k_2$).

We have noted that any solution to the problem of endogeneity requires instruments. This does not mean that valid instruments actually exist.

15.2 Reduced Form

The reduced form relationship between the variables or "regressors" $\boldsymbol{x}_i$ and the instruments $\boldsymbol{z}_i$ is found by linear projection. Let

$$\boldsymbol{\Gamma} = \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{z}_i'\right)^{-1} \mathbb{E}\left(\boldsymbol{z}_i \boldsymbol{x}_i'\right)$$

be the $\ell \times k$ matrix of coefficients from a projection of $\boldsymbol{x}_i$ on $\boldsymbol{z}_i$, and define

$$\boldsymbol{u}_i = \boldsymbol{x}_i - \boldsymbol{\Gamma}' \boldsymbol{z}_i$$

as the projection error. Then the reduced form linear relationship between $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ is

$$\boldsymbol{x}_i = \boldsymbol{\Gamma}' \boldsymbol{z}_i + \boldsymbol{u}_i. \tag{15.5}$$

In matrix notation, we can write (15.5) as

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U} \tag{15.6}$$

where $\boldsymbol{U}$ is $n \times k$.

By construction,

$$\mathbb{E}(\boldsymbol{z}_i \boldsymbol{u}_i') = \boldsymbol{0},$$

so (15.5) is a projection and can be estimated by OLS:

$$\boldsymbol{x}_i = \widehat{\boldsymbol{\Gamma}}' \boldsymbol{z}_i + \hat{\boldsymbol{u}}_i.$$

or

$$\boldsymbol{X} = \boldsymbol{Z}\widehat{\boldsymbol{\Gamma}} + \widehat{\boldsymbol{U}}$$

where

$$\widehat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \left(\boldsymbol{Z}'\boldsymbol{X}\right).$$

Substituting (15.6) into (15.2), we find

$$\begin{aligned} \boldsymbol{y} &= (\boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U})\,\boldsymbol{\beta} + \boldsymbol{e} \\ &= \boldsymbol{Z}\boldsymbol{\lambda} + \boldsymbol{v}, \end{aligned} \tag{15.7}$$

where

$$\boldsymbol{\lambda} = \boldsymbol{\Gamma}\boldsymbol{\beta} \tag{15.8}$$

and

$$\boldsymbol{v} = \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{e}.$$

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 (15.7) is a projection equation and may be estimated by OLS. This is

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Z}\hat{\boldsymbol{\lambda}} + \hat{\boldsymbol{v}}, \\ \hat{\boldsymbol{\lambda}} &= \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} \left(\boldsymbol{Z}'\boldsymbol{y}\right) \end{aligned}$$

The equation (15.7) is the reduced form for $\boldsymbol{y}$. (15.6) and (15.7) together are the **reduced form equations** for the system

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Z}\boldsymbol{\lambda} + \boldsymbol{v} \\ \boldsymbol{X} &= \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U}. \end{aligned}$$

As we showed above, OLS yields the reduced-form estimates $\left(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\Gamma}}\right)$

15.3 Identification

The structural parameter $\boldsymbol{\beta}$ relates to $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$ through (15.8). The parameter $\boldsymbol{\beta}$ is **identified**, meaning that it can be recovered from the reduced form, if

$$\operatorname{rank}(\boldsymbol{\Gamma}) = k. \tag{15.9}$$

Assume that (15.9) holds. If $\ell = k$, then $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}$. If $\ell > k$, then for any $\boldsymbol{W} > 0$, $\boldsymbol{\beta} = \left(\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\Gamma}\right)^{-1}\boldsymbol{\Gamma}'\boldsymbol{W}\boldsymbol{\lambda}$.

If (15.9) is not satisfied, then $\boldsymbol{\beta}$ cannot be recovered from $(\boldsymbol{\lambda}, \boldsymbol{\Gamma})$. Note that a necessary (although not sufficient) condition for (15.9) is $\ell \geq k$.

Since $\boldsymbol{Z}$ and $\boldsymbol{X}$ have the common variables $\boldsymbol{X}_1$, we can rewrite some of the expressions. Using (15.3) and (15.4) to make the matrix partitions $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2]$ and $\boldsymbol{X} = [\boldsymbol{Z}_1, \boldsymbol{X}_2]$, we can partition $\boldsymbol{\Gamma}$ as

$$\begin{aligned}
\boldsymbol{\Gamma} &= \begin{array}{cc} \begin{array}{cc} k_1 & k_2 \end{array} & \\ \left[\begin{array}{cc} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} \end{array}\right] & \begin{array}{c} \ell_1 \\ \ell_2 \end{array} \end{array} \\
&= \left[\begin{array}{cc} \boldsymbol{I} & \boldsymbol{\Gamma}_{12} \\ \mathbf{0} & \boldsymbol{\Gamma}_{22} \end{array}\right]
\end{aligned}$$

(15.6) can be rewritten as

$$\begin{aligned}
\boldsymbol{X}_1 &= \boldsymbol{Z}_1 \\
\boldsymbol{X}_2 &= \boldsymbol{Z}_1\boldsymbol{\Gamma}_{12} + \boldsymbol{Z}_2\boldsymbol{\Gamma}_{22} + \boldsymbol{U}_2.
\end{aligned} \tag{15.10}$$

$\boldsymbol{\beta}$ 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 (15.10).

15.4 Estimation

The model can be written as

$$\begin{aligned}
y_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \\
\mathbb{E}\left(\boldsymbol{z}_i e_i\right) &= \mathbf{0}
\end{aligned}$$

or

$$\begin{aligned}
\mathbb{E}\boldsymbol{g}_i(\boldsymbol{\beta}) &= \mathbf{0} \\
\boldsymbol{g}_i(\boldsymbol{\beta}) &= \boldsymbol{z}_i\left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right).
\end{aligned}$$

This is a moment condition model. Appropriate estimators include GMM and EL. The estimators and distribution theory developed in those Chapter 8 and 9 directly apply. Recall that the GMM estimator, for given weight matrix $\boldsymbol{W}_n$, is

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{y}.$$

15.5 Special Cases: IV and 2SLS

If the model is just-identified, so that $k = \ell$, then the formula for GMM simplifies. We find that

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{y} \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{W}_n^{-1}\left(\boldsymbol{X}'\boldsymbol{Z}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\boldsymbol{W}_n\boldsymbol{Z}'\boldsymbol{y} \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}
\end{aligned}$$

This estimator is often called the **instrumental variables estimator** (IV) of $\boldsymbol{\beta}$, where $\boldsymbol{Z}$ is used as an instrument for $\boldsymbol{X}$. Observe that the weight matrix $\boldsymbol{W}_n$ has disappeared. In the just-identified case, the weight matrix places no role. This is also the method of moments estimator of $\boldsymbol{\beta}$, and the EL estimator. Another interpretation stems from the fact that since $\boldsymbol{\beta} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\lambda}$, we can construct the **Indirect Least Squares** (ILS) estimator:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \widehat{\boldsymbol{\Gamma}}^{-1}\widehat{\boldsymbol{\lambda}} \\
&= \left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)\right)^{-1}\left(\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right)\right) \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right) \\
&= \left(\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{y}\right).
\end{aligned}$$

which again is the IV estimator.

Recall that the optimal weight matrix is an estimate of the inverse of $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'e_i^2\right)$. In the special case that $\mathbb{E}\left(e_i^2 \mid \boldsymbol{z}_i\right) = \sigma^2$ (homoskedasticity), then $\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\sigma^2 \propto \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)$ suggesting the weight matrix $\boldsymbol{W}_n = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$. Using this choice, the GMM estimator equals

$$\widehat{\boldsymbol{\beta}}_{2SLS} = \left(\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

This is called the **two-stage-least squares** (2SLS) estimator. It was originally proposed by Theil (1953) and Basmann (1957), and is the classic estimator for linear equations with instruments. Under the homoskedasticity assumption, the 2SLS estimator is efficient GMM, but otherwise it is inefficient.

It is useful to observe that writing

$$\begin{aligned}
\boldsymbol{P} &= \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}' \\
\widehat{\boldsymbol{X}} &= \boldsymbol{P}\boldsymbol{X} = \boldsymbol{Z}\widehat{\boldsymbol{\Gamma}}
\end{aligned}$$

then the 2SLS estimator is

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{y} \\
&= \left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{y}.
\end{aligned}$$

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 $\widehat{\boldsymbol{X}} = \boldsymbol{Z}\widehat{\boldsymbol{\Gamma}} = \boldsymbol{P}\boldsymbol{X}$.
- Second, regress $\boldsymbol{y}$ on $\widehat{\boldsymbol{X}}$, vis., $\widehat{\boldsymbol{\beta}} = \left(\widehat{\boldsymbol{X}}'\widehat{\boldsymbol{X}}\right)^{-1}\widehat{\boldsymbol{X}}'\boldsymbol{y}$.

It is useful to scrutinize the projection $\widehat{\boldsymbol{X}}$. Recall, $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $\boldsymbol{Z} = [\boldsymbol{X}_1, \boldsymbol{Z}_2]$. Then

$$\begin{aligned}
\widehat{\boldsymbol{X}} &= \left[\widehat{\boldsymbol{X}}_1, \widehat{\boldsymbol{X}}_2\right] \\
&= [\boldsymbol{P}\boldsymbol{X}_1, \boldsymbol{P}\boldsymbol{X}_2] \\
&= [\boldsymbol{X}_1, \boldsymbol{P}\boldsymbol{X}_2] \\
&= \left[\boldsymbol{X}_1, \widehat{\boldsymbol{X}}_2\right],
\end{aligned}$$

since $\boldsymbol{X}_1$ lies in the span of $\boldsymbol{Z}$. 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:

$$\widehat{\boldsymbol{X}}_2 = \boldsymbol{Z}_1\widehat{\boldsymbol{\Gamma}}_{12} + \boldsymbol{Z}_2\widehat{\boldsymbol{\Gamma}}_{22}.$$

15.6 Bekker Asymptotics

Bekker (1994) used an alternative asymptotic framework to analyze the finite-sample bias in the 2SLS estimator. Here we present a simplified version of one of his results. In our notation, the model is

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} && (15.11) \\ \boldsymbol{X} &= \boldsymbol{Z}\boldsymbol{\Gamma} + \boldsymbol{U} && (15.12) \\ \boldsymbol{\xi} &= (\boldsymbol{e}, \boldsymbol{U}) \\ \mathbb{E}\left(\boldsymbol{\xi} \mid \boldsymbol{Z}\right) &= \boldsymbol{0} \\ \mathbb{E}\left(\boldsymbol{\xi}'\boldsymbol{\xi} \mid \boldsymbol{Z}\right) &= \boldsymbol{S} \end{aligned}$$

As before, $\boldsymbol{Z}$ is $n \times l$ so there are l instruments.

First, let's analyze the approximate bias of OLS applied to (15.11). Using (15.12),

$$\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e}\right) = \mathbb{E}\left(\boldsymbol{x}_i e_i\right) = \boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i e_i\right) + \mathbb{E}\left(\boldsymbol{u}_i e_i\right) = \boldsymbol{s}_{21}$$

and

$$\begin{aligned} \mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right) &= \mathbb{E}\left(\boldsymbol{x}_i\boldsymbol{x}_i'\right) \\ &= \boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma} + \mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma} + \boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{u}_i'\right) + \mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{u}_i'\right) \\ &= \boldsymbol{\Gamma}'\boldsymbol{Q}\boldsymbol{\Gamma} + \boldsymbol{S}_{22} \end{aligned}$$

where $\boldsymbol{Q} = \mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)$. Hence by a first-order approximation

$$\begin{aligned} \mathbb{E}\left(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}\right) &\approx \left(\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)\right)^{-1}\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{e}\right) \\ &= \left(\boldsymbol{\Gamma}'\boldsymbol{Q}\boldsymbol{\Gamma} + \boldsymbol{S}_{22}\right)^{-1}\boldsymbol{s}_{21} && (15.13) \end{aligned}$$

which is zero only when $\boldsymbol{s}_{21} = \boldsymbol{0}$ (when $\boldsymbol{X}$ is exogenous).

We now derive a similar result for the 2SLS estimator.

$$\hat{\boldsymbol{\beta}}_{2SLS} = \left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{y}\right).$$

Let $\boldsymbol{P} = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$. By the spectral decomposition of an idempotent matrix, $\boldsymbol{P} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{\Lambda} = \text{diag}\left(\boldsymbol{I}_l, \boldsymbol{0}\right)$. Let $\boldsymbol{Q} = \boldsymbol{H}'\boldsymbol{\xi}\boldsymbol{S}^{-1/2}$ which satisfies $\mathbb{E}\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_n$ and partition $\boldsymbol{Q} = (\boldsymbol{q}_1' \ \boldsymbol{Q}_2')$ where $\boldsymbol{q}_1$ is $l \times 1$. Hence

$$\begin{aligned} \mathbb{E}\left(\frac{1}{n}\boldsymbol{\xi}'\boldsymbol{P}\boldsymbol{\xi} \mid \boldsymbol{Z}\right) &= \frac{1}{n}\boldsymbol{S}^{1/2\prime}\mathbb{E}\left(\boldsymbol{Q}'\boldsymbol{\Lambda}\boldsymbol{Q} \mid \boldsymbol{Z}\right)\boldsymbol{S}^{1/2} \\ &= \frac{1}{n}\boldsymbol{S}^{1/2\prime}\mathbb{E}\left(\frac{1}{n}\boldsymbol{q}_1'\boldsymbol{q}_1\right)\boldsymbol{S}^{1/2} \\ &= \frac{l}{n}\boldsymbol{S}^{1/2\prime}\boldsymbol{S}^{1/2} \\ &= \alpha\boldsymbol{S} \end{aligned}$$

where

$$\alpha = \frac{l}{n}.$$

Using (15.12) and this result,

$$\frac{1}{n}\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{e}\right) = \frac{1}{n}\mathbb{E}\left(\boldsymbol{\Gamma}'\boldsymbol{Z}'\boldsymbol{e}\right) + \frac{1}{n}\mathbb{E}\left(\boldsymbol{U}'\boldsymbol{P}\boldsymbol{e}\right) = \alpha\boldsymbol{s}_{21},$$

and

$$\begin{aligned}\frac{1}{n}\mathbb{E}\left(\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\right) &= \boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma} + \boldsymbol{\Gamma}'\mathbb{E}\left(\boldsymbol{z}_i\boldsymbol{u}_i\right) + \mathbb{E}\left(\boldsymbol{u}_i\boldsymbol{z}_i'\right)\boldsymbol{\Gamma} + \frac{1}{n}\mathbb{E}\left(\boldsymbol{U}'\boldsymbol{P}\boldsymbol{U}\right)\\ &= \boldsymbol{\Gamma}'\boldsymbol{Q}\boldsymbol{\Gamma} + \alpha\boldsymbol{S}_{22}.\end{aligned}$$

Together

$$\begin{aligned}\mathbb{E}\left(\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}\right) &\approx \left(\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{X}\right)\right)^{-1}\mathbb{E}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{P}\boldsymbol{e}\right)\\ &= \alpha\left(\boldsymbol{\Gamma}'\boldsymbol{Q}\boldsymbol{\Gamma} + \alpha\boldsymbol{S}_{22}\right)^{-1}\boldsymbol{s}_{21}. \qquad (15.14)\end{aligned}$$

In general this is non-zero, except when $\boldsymbol{s}_{21} = 0$ (when $\boldsymbol{X}$ is exogenous). It is also close to zero when $\alpha = 0$. Bekker (1994) pointed out that it also has the reverse implication – that when $\alpha = l/n$ is large, the bias in the 2SLS estimator will be large. Indeed as $\alpha \to 1$, the expression in (15.14) approaches that in (15.13), indicating that the bias in 2SLS approaches that of OLS as the number of instruments increases.

Bekker (1994) showed further that under the alternative asymptotic approximation that α is fixed as $n \to \infty$ (so that the number of instruments goes to infinity proportionately with sample size) then the expression in (15.14) is the probability limit of $\hat{\boldsymbol{\beta}}_{2SLS} - \boldsymbol{\beta}$

15.7 Identification Failure

Recall the reduced form equation

$$\boldsymbol{X}_2 = \boldsymbol{Z}_1\boldsymbol{\Gamma}_{12} + \boldsymbol{Z}_2\boldsymbol{\Gamma}_{22} + \boldsymbol{U}_2.$$

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 $k = l = 1$ (so there is no $\boldsymbol{Z}_1$). Then the model may be written as

$$\begin{aligned}y_i &= x_i\beta + e_i\\ x_i &= z_i\gamma + u_i\end{aligned}$$

and $\Gamma_{22} = \gamma = \mathbb{E}\left(z_i x_i\right)/\mathbb{E}z_i^2$. We see that β is identified if and only if $\gamma \neq 0$, which occurs when $\mathbb{E}\left(x_i z_i\right) \neq 0$. Thus identification hinges on the existence of correlation between the excluded exogenous variable and the included endogenous variable.

Suppose this condition fails, so $\mathbb{E}\left(x_i z_i\right) = 0$. Then by the CLT

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i e_i \xrightarrow{d} N_1 \sim \mathrm{N}\left(0, \mathbb{E}\left(z_i^2 e_i^2\right)\right) \qquad (15.15)$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i x_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i \xrightarrow{d} N_2 \sim \mathrm{N}\left(0, \mathbb{E}\left(z_i^2 u_i^2\right)\right) \qquad (15.16)$$

therefore

$$\hat{\beta} - \beta = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i e_i}{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i x_i} \xrightarrow{d} \frac{N_1}{N_2} \sim \text{Cauchy},$$

since the ratio of two normals is Cauchy. This is particularly nasty, as the Cauchy distribution does not have a finite mean. This result carries over to more general settings, and was examined by Phillips (1989) and Choi and Phillips (1992).

Suppose that identification does not completely fail, but is *weak*. This occurs when Γ_{22} is full rank, but *small*. This can be handled in an asymptotic analysis by modeling it as local-to-zero, viz

$$\boldsymbol{\Gamma}_{22} = n^{-1/2}\boldsymbol{C},$$

where $\boldsymbol{C}$ is a full rank matrix. The $n^{-1/2}$ is picked because it provides just the right balancing to allow a rich distribution theory.

To see the consequences, once again take the simple case $k = l = 1$. Here, the instrument x_i is weak for z_i if

$$\gamma = n^{-1/2}c.$$

Then (15.15) is unaffected, but (15.16) instead takes the form

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i x_i &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i^2\gamma + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i \\ &= \frac{1}{n}\sum_{i=1}^{n} z_i^2 c + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i \\ &\xrightarrow{d} Qc + N_2\end{aligned}$$

therefore

$$\hat{\beta} - \beta \xrightarrow{d} \frac{N_1}{Qc + N_2}.$$

As in the case of complete identification failure, we find that $\hat{\beta}$ is inconsistent for β and the asymptotic distribution of $\hat{\beta}$ is non-normal. In addition, standard test statistics have non-standard distributions, meaning that inferences about parameters of interest can be misleading.

The distribution theory for this model was developed by Staiger and Stock (1997) and extended to nonlinear GMM estimation by Stock and Wright (2000). Further results on testing were obtained by Wang and Zivot (1998).

The bottom line is that it is highly desirable to avoid identification failure. Once again, the equation to focus on is the reduced form

$$\boldsymbol{X}_2 = \boldsymbol{Z}_1\boldsymbol{\Gamma}_{12} + \boldsymbol{Z}_2\boldsymbol{\Gamma}_{22} + \boldsymbol{U}_2$$

and identification requires rank$(\boldsymbol{\Gamma}_{22}) = k_2$. If $k_2 = 1$, this requires $\Gamma_{22} \neq \boldsymbol{0}$, which is straightforward to assess using a hypothesis test on the reduced form. Therefore in the case of $k_2 = 1$ (one RHS endogenous variable), one constructive recommendation is to explicitly estimate the reduced form equation for $\boldsymbol{X}_2$, construct the test of $\boldsymbol{\Gamma}_{22} = \boldsymbol{0}$, and at a minimum check that the test rejects $\mathbb{H}_0 : \Gamma_{22} = \boldsymbol{0}$.

When $k_2 > 1$, $\boldsymbol{\Gamma}_{22} \neq 0$ is not sufficient for identification. It is not even sufficient that each column of $\boldsymbol{\Gamma}_{22}$ is non-zero (each column corresponds to a distinct endogenous variable in $\boldsymbol{Z}_2$). So while a minimal check is to test that each columns of $\boldsymbol{\Gamma}_{22}$ is non-zero, this cannot be interpreted as definitive proof that $\boldsymbol{\Gamma}_{22}$ has full rank. Unfortunately, tests of deficient rank are difficult to implement. In any event, it appears reasonable to explicitly estimate and report the reduced form equations for $\boldsymbol{Z}_2$, and attempt to assess the likelihood that $\boldsymbol{\Gamma}_{22}$ has deficient rank.

Exercises

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 $\hat{\beta}$ 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.

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 $\boldsymbol{z}_i$.)

3. Take the linear model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e}.$$

Let the OLS estimator for $\boldsymbol{\beta}$ be $\hat{\boldsymbol{\beta}}$ and the OLS residual be $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

Let the IV estimator for $\boldsymbol{\beta}$ using some instrument $\boldsymbol{Z}$ be $\tilde{\boldsymbol{\beta}}$ and the IV residual be $\tilde{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}$. If $\boldsymbol{X}$ is indeed endogeneous, will IV "fit" better than OLS, in the sense that $\tilde{\boldsymbol{e}}'\tilde{\boldsymbol{e}} < \hat{\boldsymbol{e}}'\hat{\boldsymbol{e}}$, at least in large samples?

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

$$\boldsymbol{X} = \boldsymbol{Z\Gamma} + \boldsymbol{U}$$

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 $\hat{\boldsymbol{\Gamma}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{X}\right)$.

5. In the structural model

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\beta} + \boldsymbol{e} \\ \boldsymbol{X} &= \boldsymbol{Z\Gamma} + \boldsymbol{U} \end{aligned}$$

with $\boldsymbol{\Gamma}$ $l \times k$, $l \geq k$, we claim that $\boldsymbol{\beta}$ is identified (can be recovered from the reduced form) if $\text{rank}(\boldsymbol{\Gamma}) = k$. Explain why this is true. That is, show that if $\text{rank}(\boldsymbol{\Gamma}) < k$ then $\boldsymbol{\beta}$ cannot be identified.

6. 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?

(c) Find the efficient GMM estimator of β based on the moment condition

$$\mathbb{E}\left(\boldsymbol{z}_i\left(y_i - x_i\beta\right)\right) = \mathbf{0}.$$

Does this differ from 2SLS and/or OLS?

7. 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?

8. The data file `card.dat` is taken from Card (1995). There are 2215 observations with 29 variables, listed in card.pdf. We want to estimate a wage equation

$$\log(Wage) = \beta_0 + \beta_1 Educ + \beta_2 Exper + \beta_3 Exper^2 + \beta_4 South + \beta_5 Black + \boldsymbol{e}$$

where $Educ = Eduation$ (Years) $Exper = Experience$ (Years), and $South$ and $Black$ are regional and racial dummy variables.

(a) Estimate the model by OLS. Report estimates and standard errors.

(b) Now treat $Education$ as endogenous, and the remaining variables as exogenous. Estimate the model by 2SLS, using the instrument $near4$, a dummy indicating that the observation lives near a 4-year college. Report estimates and standard errors.

(c) Re-estimate by 2SLS (report estimates and standard errors) adding three additional instruments: *near2* (a dummy indicating that the observation lives near a 2-year college), $fatheduc$ (the education, in years, of the father) and $motheduc$ (the education, in years, of the mother).

(d) Re-estimate the model 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.

(e) Calculate and report the J statistic for overidentification.

(f) Discuss your findings.

Chapter 16

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).

16.1 Stationarity and Ergodicity

Definition 16.1.1 $\{y_t\}$ *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 16.1.2 $\{y_t\}$ *is* ***strictly stationary*** *if the joint distribution of* $(y_t, ..., y_{t-k})$ *is independent of* t *for all* k.

Definition 16.1.3 *A stationary time series is* ***ergodic*** *if* $\gamma(k) \rightarrow 0$ *as* $k \rightarrow \infty$.

The following two theorems are essential to the analysis of stationary time series. The proofs are rather difficult, however.

Theorem 16.1.1 *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 16.1.2 *(Ergodic Theorem). If* y_t *is strictly stationary and ergodic and* $\mathbb{E}\left|y_t\right| < \infty$*, 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:

$$\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} y_t$$

The sample autocovariance

$$\hat{\gamma}(k) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{\mu}\right)\left(y_{t-k} - \hat{\mu}\right).$$

The sample autocorrelation

$$\hat{\rho}(k) = \frac{\hat{\gamma}(k)}{\hat{\gamma}(0)}.$$

Theorem 16.1.3 *If* y_t *is strictly stationary and ergodic and* $\mathbb{E}y_t^2 < \infty$*, then as* $T \to \infty$*,*

1. $\hat{\mu} \xrightarrow{p} \mathbb{E}(y_t)$;
2. $\hat{\gamma}(k) \xrightarrow{p} \gamma(k)$;
3. $\hat{\rho}(k) \xrightarrow{p} \rho(k)$.

Proof of Theorem 16.1.3. Part (1) is a direct consequence of the Ergodic theorem. For Part (2), note that

$$\begin{aligned}\hat{\gamma}(k) &= \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \hat{\mu}\right)\left(y_{t-k} - \hat{\mu}\right) \\ &= \frac{1}{T}\sum_{t=1}^{T} y_t y_{t-k} - \frac{1}{T}\sum_{t=1}^{T} y_t\hat{\mu} - \frac{1}{T}\sum_{t=1}^{T} y_{t-k}\hat{\mu} + \hat{\mu}^2.\end{aligned}$$

By Theorem 16.1.1 above, the sequence $y_t y_{t-k}$ is strictly stationary and ergodic, and it has a finite mean by the assumption that $\mathbb{E}y_t^2 < \infty$. 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

$$\hat{\gamma}(k) \xrightarrow{p} \mathbb{E}(y_t y_{t-k}) - \mu^2 - \mu^2 + \mu^2 = \mathbb{E}(y_t y_{t-k}) - \mu^2 = \gamma(k).$$

Part (3) follows by the continuous mapping theorem: $\hat{\rho}(k) = \hat{\gamma}(k)/\hat{\gamma}(0) \xrightarrow{p} \gamma(k)/\gamma(0) = \rho(k)$.

16.2 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 16.2.1 *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 $k > 0$, $\mathbb{E}\left(y_{t-k} e_t\right) = 0$.

16.3 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 $\mathbb{E}e_t^2 = \sigma^2 < \infty$.

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 16.3.1 *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:

$$\mathbb{E}y_t = \sum_{k=0}^{\infty} \alpha^k \mathbb{E}\left(e_{t-k}\right) = 0$$

$$\operatorname{var}(y_t) = \sum_{k=0}^{\infty} \alpha^{2k} \operatorname{var}\left(e_{t-k}\right) = \frac{\sigma^2}{1-\alpha^2}.$$

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 $\mathbb{E}y_t = \alpha_0/(1-\alpha_1)$.

16.4 Lag Operator

An algebraic construct which is useful for the analysis of autoregressive models is the lag operator.

Definition 16.4.1 *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

$$\left(1 - \alpha \mathrm{L}\right) y_t = e_t.$$

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 16.3.1, 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).

16.5 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 16.5.1 *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.

16.6 Estimation

Let

$$\begin{aligned} \boldsymbol{x}_t &= \begin{pmatrix} 1 & y_{t-1} & y_{t-2} & \cdots & y_{t-k} \end{pmatrix}' \\ \boldsymbol{\beta} &= \begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \cdots & \alpha_k \end{pmatrix}'. \end{aligned}$$

Then the model can be written as

$$y_t = \boldsymbol{x}_t' \boldsymbol{\beta} + e_t.$$

The OLS estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}' \boldsymbol{y}.$$

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 16.1.1, it is also strictly stationary and ergodic. Thus

$$\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t e_t = \frac{1}{T}\sum_{t=1}^{T} u_t \xrightarrow{p} \mathbb{E}\left(u_t\right) = 0. \tag{16.1}$$

The vector $\boldsymbol{x}_t$ is strictly stationary and ergodic, and by Theorem 16.1.1, 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 (16.1) 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 16.6.1 *If the AR(k) process y_t is strictly stationary and ergodic and $\mathbb{E}y_t^2 < \infty$, then $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$ as $T \to \infty$.*

16.7 Asymptotic Distribution

Theorem 16.7.1 *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 16.7.1 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 16.7.2 *If the AR(k) process y_t is strictly stationary and ergodic and $\mathbb{E}y_t^4 < \infty$, 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.

16.8 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 $\hat{\boldsymbol{\beta}}$ and residuals $\hat{e}_t$.
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 $\{\hat{e}_1, ..., \hat{e}_T\}$.
4. Create the bootstrap series y_t^* by the recursive formula

$$y_t^* = \hat{\alpha}_0 + \hat{\alpha}_1 y_{t-1}^* + \hat{\alpha}_2 y_{t-2}^* + \cdots + \hat{\alpha}_k y_{t-k}^* + e_t^*.$$

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.

16.9 Trend Stationarity

$$y_t = \mu_0 + \mu_1 t + S_t \tag{16.2}$$
$$S_t = \rho_1 S_{t-1} + \rho_2 S_{t-2} + \cdots + \rho_k S_{t-k} + e_t, \tag{16.3}$$

or

$$y_t = \alpha_0 + \alpha_1 t + \rho_1 y_{t-1} + \rho_2 y_{t-1} + \cdots + \rho_k y_{t-k} + e_t. \tag{16.4}$$

There are two essentially equivalent ways to estimate the autoregressive parameters $(\alpha_1, ..., \alpha_k)$.

- You can estimate (16.4) by OLS.
- You can estimate (16.2)-(16.3) sequentially by OLS. That is, first estimate (16.2), get the residual $\hat{S}_t$, and then perform regression (16.3) 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.

16.10 Testing for Omitted Serial Correlation

For simplicity, let the null hypothesis be an AR(1):

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + u_t. \tag{16.5}$$

We are interested in the question if the error u_t is serially correlated. We model this as an AR(1):

$$u_t = \theta u_{t-1} + e_t \tag{16.6}$$

with e_t a MDS. The hypothesis of no omitted serial correlation is

$$\begin{aligned}\mathbb{H}_0 &: \theta = 0\\ \mathbb{H}_1 &: \theta \neq 0.\end{aligned}$$

We want to test $\mathbb{H}_0$ against $\mathbb{H}_1$.

To combine (16.5) and (16.6), we take (16.5) 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 (16.5), 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.

16.11 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

$$AIC(k) = \log \hat{\sigma}^2(k) + \frac{2k}{T},$$

where $\hat{\sigma}^2(k)$ 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.

16.12 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:

$$\Delta y_t = \rho_0 y_{t-1} + \rho_1 \Delta y_{t-1} + \cdots + \rho_{k-1} \Delta y_{t-(k-1)} + e_t. \tag{16.7}$$

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 (16.7) 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 (16.7). 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 16.12.1 ***Dickey-Fuller Theorem.***
If $\rho_0 = 0$ then as $T \to \infty$,

$$T\hat{\rho}_0 \xrightarrow{d} (1 - \rho_1 - \rho_2 - \cdots - \rho_{k-1})\, DF_\alpha$$

$$ADF = \frac{\hat{\rho}_0}{s(\hat{\rho}_0)} \to DF_t.$$

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 $\hat{\rho}_0$ 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 $\hat{\rho}_0$ 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

$$\Delta y_t = \mu_1 + \mu_2 t + \rho_0 y_{t-1} + \rho_1 \Delta y_{t-1} + \cdots + \rho_{k-1} \Delta y_{t-(k-1)} + e_t. \tag{16.8}$$

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 (16.8).

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 17

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.

17.1 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).

17.2 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

$$\hat{\boldsymbol{a}}_j = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}_j.$$

An alternative way to compute this is as follows. Note that

$$\hat{\boldsymbol{a}}_j' = \boldsymbol{y}_j'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

And if we stack these to create the estimate $\hat{A}$, we find

$$\begin{aligned}\hat{\boldsymbol{A}} &= \begin{pmatrix} \boldsymbol{y}_1' \\ \boldsymbol{y}_2' \\ \vdots \\ \boldsymbol{y}_{m+1}' \end{pmatrix} \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\ &= \boldsymbol{Y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1},\end{aligned}$$

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)

17.3 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.

17.4 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

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t, \tag{17.1}$$

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.

17.5 Testing for Omitted Serial Correlation

Consider the problem of testing for omitted serial correlation in equation (17.1). Suppose that e_t is an AR(1). Then

$$\begin{aligned} y_t &= \boldsymbol{x}_t'\boldsymbol{\beta} + e_t \\ e_t &= \theta e_{t-1} + u_t \\ \mathbb{E}\left(u_t \mid \mathcal{F}_{t-1}\right) &= 0. \end{aligned} \tag{17.2}$$

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

$$y_t = \theta y_{t-1} + \boldsymbol{x}_t'\boldsymbol{\beta} + \boldsymbol{x}_{t-1}'\boldsymbol{\gamma} + u_t, \tag{17.3}$$

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 (17.2) is a special case of (17.3), 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 (17.2) 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 (17.2) is uncommon in recent applications.

17.6 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

$$\begin{aligned} AIC(k) &= \log\det\left(\hat{\boldsymbol{\Omega}}(k)\right) + 2\frac{p}{T} \\ BIC(k) &= \log\det\left(\hat{\boldsymbol{\Omega}}(k)\right) + \frac{p\log(T)}{T} \\ \hat{\boldsymbol{\Omega}}(k) &= \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{e}}_t(k)\hat{\boldsymbol{e}}_t(k)' \\ p &= m(km+1) \end{aligned}$$

where p is the number of parameters in the model, and $\hat{\boldsymbol{e}}_t(k)$ is the OLS residual vector from the model with k lags. The log determinant is the criterion from the multivariate normal likelihood.

17.7 Granger Causality

Partition the data vector into $(\boldsymbol{y}_t, \boldsymbol{z}_t)$. Define the two information sets

$$\begin{aligned}\mathcal{F}_{1t} &= \left(\boldsymbol{y}_t, \boldsymbol{y}_{t-1}, \boldsymbol{y}_{t-2}, ...\right)\\ \mathcal{F}_{2t} &= \left(\boldsymbol{y}_t, \boldsymbol{z}_t, \boldsymbol{y}_{t-1}, \boldsymbol{z}_{t-1}, \boldsymbol{y}_{t-2}, \boldsymbol{z}_{t-2}, , ...\right)\end{aligned}$$

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

$$\boldsymbol{y}_t = \alpha + \rho_1 \boldsymbol{y}_{t-1} + \cdots + \rho_k \boldsymbol{y}_{t-k} + \boldsymbol{z}_{t-1}' \gamma_1 + \cdots + \boldsymbol{z}_{t-k}' \gamma_k + e_t.$$

In this equation, $\boldsymbol{z}_t$ does not Granger-cause $\boldsymbol{y}_t$ if and only if

$$\mathbb{H}_0 : \gamma_1 = \gamma_2 = \cdots = \gamma_k = 0.$$

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.

17.8 Cointegration

The idea of cointegration is due to Granger (1981), and was articulated in detail by Engle and Granger (1987).

Definition 17.8.1 *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 $\hat{\beta}_2 = (\boldsymbol{y}_2'\boldsymbol{y}_2)^{-1}\boldsymbol{y}_2'\boldsymbol{y}_1 \xrightarrow{p} \beta_2$. Furthermore this estimation is super-consistent: $T(\hat{\beta}_2 - \beta_2) \xrightarrow{d} Limit$, as first shown by Stock (1987). This is not, in general, a good method to estimate $\boldsymbol{\beta}$, but 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 $\hat{z}_t = \hat{\boldsymbol{\beta}}'\boldsymbol{y}_t$, from OLS of y_{1t} on y_{2t}. Their justification was Stock's result that $\hat{\boldsymbol{\beta}}$ is super-consistent under $\mathbb{H}_1$. Under $\mathbb{H}_0$, however, $\hat{\boldsymbol{\beta}}$ 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).

17.9 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 17.9.1 ***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 $\mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$.

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 18

Limited Dependent Variables

A "limited dependent variable" y is one which takes a "limited" set of values. 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.

18.1 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: $F(u) = \left(1 + e^{-u}\right)^{-1}$.

- 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 $\hat{\boldsymbol{\beta}}$ 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.

18.2 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

$$\mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

is the conditional mean, this motivates the label Poisson "regression."

Also observe that the model implies that

$$\operatorname{var}\left(y_i \mid \boldsymbol{x}_i\right) = \lambda_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}),$$

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.

18.3 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

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0 \end{cases}. \tag{18.1}$$

(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 (18.1). 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 $\mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right)$, not $\mathbb{E}\left(y_i^* \mid \boldsymbol{x}_i\right) = \boldsymbol{x}_i'\boldsymbol{\beta}$, 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 $\mathbb{E}\left(y_i \mid \boldsymbol{x}_i\right)$ 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 $\hat{\boldsymbol{\beta}}$ which maximizes $\log L(\boldsymbol{\beta})$.

18.4 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

$$\left(\begin{array}{c} e_{0i} \\ e_{1i} \end{array}\right) \sim \mathrm{N}\left(0, \left(\begin{array}{cc} 1 & \rho \\ \rho & \sigma^2 \end{array}\right)\right).$$

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

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \rho \lambda\left(\boldsymbol{z}_i'\boldsymbol{\gamma}\right) + u_i \tag{18.2}$$

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 $\hat{\boldsymbol{\gamma}}$ from a Probit, using regressors $\boldsymbol{z}_i$. The binary dependent variable is T_i.
- Estimate $\left(\hat{\boldsymbol{\beta}}, \hat{\rho}\right)$ from OLS of y_i on $\boldsymbol{x}_i$ and $\lambda(\boldsymbol{z}_i'\hat{\boldsymbol{\gamma}})$.
- 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 $\lambda\left(\boldsymbol{z}_i'\hat{\gamma}\right)$ 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 $\lambda\left(\boldsymbol{z}_i'\hat{\gamma}\right)$ 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 $\lambda\left(\boldsymbol{z}_i'\hat{\gamma}\right)$ is not collinear with $\boldsymbol{x}_i$, and hence improve the second stage estimator's precision.

Chapter 19

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*:

$$\{y_{it}, \boldsymbol{x}_{it}\} : \text{For } i = 1, ..., n, \quad t = \underline{t}_i, ..., \overline{t}_i.$$

19.1 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

$$\mathbb{E}\left(\boldsymbol{x}_{it}u_i\right) = 0 \tag{19.1}$$

If this condition fails, then OLS is inconsistent. (19.1) 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 (19.1) is true, however, OLS can be improved upon via a GLS technique. In either event, OLS appears a poor estimation choice.

Condition (19.1) is called the *random effects hypothesis.* It is a strong assumption, and most applied researchers try to avoid its use.

19.2 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

$$y_{it} = \boldsymbol{x}_{it}' \boldsymbol{\beta} + \boldsymbol{d}_i' \boldsymbol{u} + e_{it}. \tag{19.2}$$

Observe that

$$\mathbb{E}\left(e_{it} \mid \boldsymbol{x}_{it}, \boldsymbol{d}_i\right) = 0,$$

so (19.2) is a valid regression, with $\boldsymbol{d}_i$ as a regressor along with $\boldsymbol{x}_i$.

OLS on (19.2) yields estimator $\left(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{u}}\right)$. 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,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{D}}\right)\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{D}}\right)\boldsymbol{y}\right) \\ &= \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*\right)^{-1}\left(\boldsymbol{X}^{*\prime}\boldsymbol{y}^*\right), \end{aligned}$$

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 $\hat{\boldsymbol{\beta}}$ 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.

19.3 Dynamic Panel Regression

A dynamic panel regression has a lagged dependent variable

$$y_{it} = \alpha y_{it-1} + \boldsymbol{x}_{it}'\boldsymbol{\beta} + u_i + e_{it}. \tag{19.3}$$

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 (19.3) eliminates the individual-specific effect:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta \boldsymbol{x}_{it}'\boldsymbol{\beta} + \Delta e_{it}. \tag{19.4}$$

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 (19.4) 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 (19.4) 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.

Chapter 20

Nonparametric Density Estimation

20.1 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 $\hat{F}(x) = n^{-1}\sum_{i=1}^{n} 1(X_i \leq x)$, we cannot define $\frac{d}{dx}\hat{F}(x)$ since $\hat{F}(x)$ 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.

Definition 20.1.1 *$K(u)$ is a **second-order kernel function** if it is a symmetric zero-mean density function.*

Three common choices for kernels include the **Normal**

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$$

the **Epanechnikov**

$$K(u) = \begin{cases} \frac{3}{4}\left(1-u^2\right), & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

and the **Biweight** or **Quartic**

$$K(u) = \begin{cases} \frac{15}{16}\left(1-u^2\right)^2, & |u| \leq 1 \\ 0 & |u| > 1 \end{cases}$$

In practice, the choice between these three rarely makes a meaningful difference in the estimates.

The kernel functions are used to smooth the data. The amount of smoothing is controlled by the **bandwidth** $h > 0$. Let

$$K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right).$$

be the kernel K rescaled by the bandwidth h. The kernel density estimator of $f(x)$ is

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(X_i - x\right).$$

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 $\hat{f}(x)$ is small. The bandwidth h controls the meaning of "near".

Interestingly, $\hat{f}(x)$ is a valid density. That is, $\hat{f}(x) \geq 0$ for all x, and

$$\begin{aligned}\int_{-\infty}^{\infty} \hat{f}(x)dx &= \int_{-\infty}^{\infty} \frac{1}{n}\sum_{i=1}^{n} K_h\left(X_i - x\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K_h\left(X_i - x\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} K\left(u\right)du = 1\end{aligned}$$

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 $\hat{f}(x)$. The mean is

$$\begin{aligned}\int_{-\infty}^{\infty} x\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} xK_h\left(X_i - x\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(X_i - uh\right)K\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i\int_{-\infty}^{\infty} K\left(u\right)du + \frac{1}{n}\sum_{i=1}^{n} h\int_{-\infty}^{\infty} uK\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i\end{aligned}$$

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

$$\begin{aligned}\int_{-\infty}^{\infty} x^2\hat{f}(x)dx &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty} x^2K_h\left(X_i - x\right)dx \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(X_i - uh\right)^2K\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{2}{n}\sum_{i=1}^{n} X_ih\int_{-\infty}^{\infty} uK(u)du + \frac{1}{n}\sum_{i=1}^{n} h^2\int_{-\infty}^{\infty} u^2K\left(u\right)du \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\sigma_K^2\end{aligned}$$

where

$$\sigma_K^2 = \int_{-\infty}^{\infty} u^2K\left(u\right)du$$

is the variance of the kernel. It follows that the variance of the density $\hat{f}(x)$ is

$$\begin{aligned}\int_{-\infty}^{\infty} x^2\hat{f}(x)dx - \left(\int_{-\infty}^{\infty} x\hat{f}(x)dx\right)^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 + h^2\sigma_K^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\ &= \hat{\sigma}^2 + h^2\sigma_K^2\end{aligned}$$

Thus the variance of the estimated density is inflated by the factor $h^2\sigma_K^2$ relative to the sample moment.

20.2 Asymptotic MSE for Kernel Estimates

For fixed x and bandwidth h observe that

$$\begin{aligned}\mathbb{E}K_h(X-x) &= \int_{-\infty}^{\infty} K_h(z-x)\, f(z)dz \\ &= \int_{-\infty}^{\infty} K_h(uh)\, f(x+hu)hdu \\ &= \int_{-\infty}^{\infty} K(u)\, f(x+hu)du\end{aligned}$$

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 $\hat{f}(x)$ is then

$$Bias(x)=\mathbb{E}\hat{f}(x)-f(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}K_h(X_i-x)-f(x)=\frac{1}{2}f''(x)h^2\sigma_K^2.$$

We see that the bias of $\hat{f}(x)$ at x depends on the second derivative $f''(x)$. The sharper the derivative, the greater the bias. Intuitively, the estimator $\hat{f}(x)$ 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 $\hat{f}(x)$. 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 $(nh)^{-1}$

$$\begin{aligned}\operatorname{var}\left(\hat{f}(x)\right) &= \frac{1}{n}\operatorname{var}(K_h(X_i-x)) \\ &= \frac{1}{n}\mathbb{E}K_h(X_i-x)^2-\frac{1}{n}\left(\mathbb{E}K_h(X_i-x)\right)^2 \\ &\simeq \frac{1}{nh^2}\int_{-\infty}^{\infty}K\left(\frac{z-x}{h}\right)^2 f(z)dz-\frac{1}{n}f(x)^2 \\ &= \frac{1}{nh}\int_{-\infty}^{\infty}K(u)^2 f(x+hu)\,du \\ &\simeq \frac{f(x)}{nh}\int_{-\infty}^{\infty}K(u)^2\,du \\ &= \frac{f(x)\,R(K)}{nh}.\end{aligned}$$

where $R(K) = \int_{-\infty}^{\infty} K(u)^2 \, du$ is called the **roughness** of K.

Together, the asymptotic mean-squared error (AMSE) for fixed x is the sum of the approximate squared bias and approximate variance

$$AMSE_h(x) = \frac{1}{4} f''(x)^2 h^4 \sigma_K^4 + \frac{f(x) R(K)}{nh}.$$

A global measure of precision is the asymptotic mean integrated squared error (AMISE)

$$AMISE_h = \int AMSE_h(x) dx = \frac{h^4 \sigma_K^4 R(f'')}{4} + \frac{R(K)}{nh}. \tag{20.1}$$

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 (20.1) 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

$$h_0 = \left(\frac{R(K)}{\sigma_K^4 R(f'')} \right)^{1/5} n^{-1/5}. \tag{20.2}$$

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 and continuing literature on the subject. The asymptotically optimal choice given in (20.2) depends on $R(K)$, σ_K^2, and $R(f'')$. The first two are determined by the kernel function. Their values for the three functions introduced in the previous section are given here.

K	$\sigma_K^2 = \int_{-\infty}^{\infty} u^2 K(u) \, du$	$R(K) = \int_{-\infty}^{\infty} K(u)^2 \, du$
Gaussian	1	$1/(2\sqrt{\pi})$
Epanechnikov	1/5	1/5
Biweight	1/7	5/7

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 (20.2) but replaces $R(f'')$ with $\hat{\sigma}^{-5} R(\phi'')$, where ϕ is the N(0,1) distribution and $\hat{\sigma}^2$ is an estimate of $\sigma^2 = \text{var}(X)$. 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 $R(\phi'') = 3/(8\sqrt{\pi})$. Together with the above table, we find the reference rules for the three kernel functions introduced earlier.

Gaussian Kernel: $h_{rule} = 1.06 \hat{\sigma} n^{-1/5}$

Epanechnikov Kernel: $h_{rule} = 2.34 \hat{\sigma} n^{-1/5}$

Biweight (Quartic) Kernel: $h_{rule} = 2.78 \hat{\sigma} n^{-1/5}$

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 $\hat{f}(x)$. 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 (20.2). 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. Fortunately there are remedies, which are known as **smoothed cross-validation** which is a close cousin of the **bootstrap**.

Appendix A

Matrix Algebra

A.1 Notation

A **scalar** a is a single real number.

A **vector** $\boldsymbol{a}$ is a $k \times 1$ list of real 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, denoted $\boldsymbol{A}'$, $\boldsymbol{A}^\top$, or $\boldsymbol{A}^t$, is obtained by flipping the matrix on its diagonal. 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 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.3 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

$$\boldsymbol{a}'\boldsymbol{b} = a_1 b_1 + a_2 b_2 + \cdots + a_k b_k = \sum_{j=1}^{k} a_j b_j.$$

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

$$\begin{aligned}\boldsymbol{AB} &= \begin{bmatrix} \boldsymbol{a}_1' \\ \boldsymbol{a}_2' \\ \vdots \\ \boldsymbol{a}_k' \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_s \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1'\boldsymbol{b}_s \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2'\boldsymbol{b}_s \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_k'\boldsymbol{b}_1 & \boldsymbol{a}_k'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_k'\boldsymbol{b}_s \end{bmatrix}.\end{aligned}$$

Matrix multiplication is not commutative: in general $\boldsymbol{AB} \neq \boldsymbol{BA}.$ However, it is associative and distributive:

$$\begin{aligned}\boldsymbol{A}(\boldsymbol{BC}) &= (\boldsymbol{AB})\,\boldsymbol{C} \\ \boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) &= \boldsymbol{AB} + \boldsymbol{AC}\end{aligned}$$

An alternative way to write the matrix product is to use matrix partitions. For example,

$$\begin{aligned}\boldsymbol{AB} &= \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{A}_{11}\boldsymbol{B}_{11} + \boldsymbol{A}_{12}\boldsymbol{B}_{21} & \boldsymbol{A}_{11}\boldsymbol{B}_{12} + \boldsymbol{A}_{12}\boldsymbol{B}_{22} \\ \boldsymbol{A}_{21}\boldsymbol{B}_{11} + \boldsymbol{A}_{22}\boldsymbol{B}_{21} & \boldsymbol{A}_{21}\boldsymbol{B}_{12} + \boldsymbol{A}_{22}\boldsymbol{B}_{22} \end{bmatrix}.\end{aligned}$$

As another example,

$$\begin{aligned}\boldsymbol{AB} &= \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{A}_2 & \cdots & \boldsymbol{A}_r \end{bmatrix} \begin{bmatrix} \boldsymbol{B}_1 \\ \boldsymbol{B}_2 \\ \vdots \\ \boldsymbol{B}_r \end{bmatrix} \\ &= \boldsymbol{A}_1\boldsymbol{B}_1 + \boldsymbol{A}_2\boldsymbol{B}_2 + \cdots + \boldsymbol{A}_r\boldsymbol{B}_r \\ &= \sum_{j=1}^{r} \boldsymbol{A}_j\boldsymbol{B}_j\end{aligned}$$

An important property of the identity matrix is that if $\boldsymbol{A}$ is $k \times r$, then $\boldsymbol{AI}_r = \boldsymbol{A}$ and $\boldsymbol{I}_k\boldsymbol{A} = \boldsymbol{A}$. The $k \times r$ matrix $\boldsymbol{A}$, $r \leq k$, is called **orthogonal** if $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{I}_r$.

A.4 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

$$\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA}). \tag{A.1}$$

Indeed,

$$\begin{aligned} \operatorname{tr}(\boldsymbol{AB}) &= \operatorname{tr}\begin{bmatrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1'\boldsymbol{b}_k \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2'\boldsymbol{b}_k \\ \vdots & \vdots & & \vdots \\ \boldsymbol{a}_k'\boldsymbol{b}_1 & \boldsymbol{a}_k'\boldsymbol{b}_2 & \cdots & \boldsymbol{a}_k'\boldsymbol{b}_k \end{bmatrix} \\ &= \sum_{i=1}^{k} \boldsymbol{a}_i'\boldsymbol{b}_i \\ &= \sum_{i=1}^{k} \boldsymbol{b}_i'\boldsymbol{a}_i \\ &= \operatorname{tr}(\boldsymbol{BA}). \end{aligned}$$

A.5 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 $\boldsymbol{Ac} = \boldsymbol{0}$.

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

$$\boldsymbol{AA}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k.$$

For non-singular $\boldsymbol{A}$ and $\boldsymbol{C}$, some important properties include

$$\begin{aligned} \boldsymbol{AA}^{-1} &= \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k \\ \left(\boldsymbol{A}^{-1}\right)' &= \left(\boldsymbol{A}'\right)^{-1} \\ (\boldsymbol{AC})^{-1} &= \boldsymbol{C}^{-1}\boldsymbol{A}^{-1} \\ (\boldsymbol{A}+\boldsymbol{C})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{C}^{-1} \\ \boldsymbol{A}^{-1} - (\boldsymbol{A}+\boldsymbol{C})^{-1} &= \boldsymbol{A}^{-1}\left(\boldsymbol{A}^{-1}+\boldsymbol{C}^{-1}\right)^{-1}\boldsymbol{A}^{-1} \end{aligned}$$

Also, if $\boldsymbol{A}$ is an orthogonal matrix, then $\boldsymbol{A}^{-1} = \boldsymbol{A}'$.

Another useful result for non-singular $\boldsymbol{A}$ is known as the **Woodbury matrix identity**

$$(\boldsymbol{A}+\boldsymbol{BCD})^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\boldsymbol{BC}\left(\boldsymbol{C}+\boldsymbol{CDA}^{-1}\boldsymbol{BC}\right)^{-1}\boldsymbol{CDA}^{-1}. \tag{A.2}$$

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**

$$\left(\boldsymbol{A}-\boldsymbol{bb}'\right)^{-1}=\boldsymbol{A}^{-1}+\left(1-\boldsymbol{b}'\boldsymbol{A}^{-1}\boldsymbol{b}\right)^{-1}\boldsymbol{A}^{-1}\boldsymbol{bb}'\boldsymbol{A}^{-1}. \tag{A.3}$$

The following fact about inverting partitioned matrices is quite useful.

$$\left[\begin{array}{cc}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{array}\right]^{-1}=\left[\begin{array}{cc}\boldsymbol{A}^{11} & \boldsymbol{A}^{12}\\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22}\end{array}\right]=\left[\begin{array}{cc}\boldsymbol{A}_{11\cdot 2}^{-1} & -\boldsymbol{A}_{11\cdot 2}^{-1}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\\ -\boldsymbol{A}_{22\cdot 1}^{-1}\boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{A}_{22\cdot 1}^{-1}\end{array}\right] \tag{A.4}$$

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

$$\begin{gathered}
\boldsymbol{AA}^{-}\boldsymbol{A}=\boldsymbol{A}\\
\boldsymbol{A}^{-}\boldsymbol{AA}^{-}=\boldsymbol{A}^{-}\\
\boldsymbol{AA}^{-}\text{ is symmetric}\\
\boldsymbol{A}^{-}\boldsymbol{A}\text{ is symmetric}
\end{gathered}$$

For any matrix $\boldsymbol{A}$, the Moore-Penrose generalized inverse $\boldsymbol{A}^{-}$ exists and is unique.

For example, if

$$\boldsymbol{A}=\left[\begin{array}{cc}\boldsymbol{A}_{11} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{array}\right]$$

and when $\boldsymbol{A}_{11}^{-1}$ exists then

$$\boldsymbol{A}^{-}=\left[\begin{array}{cc}\boldsymbol{A}_{11}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{array}\right].$$

A.6 Determinant

The **determinant** is a measure of the volume of a square matrix.

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 general $k\times k$ 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

$$\begin{aligned} \det \boldsymbol{A} &= \varepsilon_{(1,2)} a_{11} a_{22} + \varepsilon_{(2,1)} a_{21} a_{12} \\ &= a_{11} a_{22} - a_{12} a_{21}. \end{aligned}$$

Some properties include

- $\det(\boldsymbol{A}) = \det(\boldsymbol{A}')$
- $\det(c\boldsymbol{A}) = c^k \det \boldsymbol{A}$
- $\det(\boldsymbol{AB}) = (\det \boldsymbol{A})(\det \boldsymbol{B})$
- $\det\left(\boldsymbol{A}^{-1}\right) = (\det \boldsymbol{A})^{-1}$
- $\det \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = (\det \boldsymbol{D}) \det\left(\boldsymbol{A} - \boldsymbol{BD}^{-1}\boldsymbol{C}\right)$ if $\det \boldsymbol{D} \neq 0$
- $\det \boldsymbol{A} \neq 0$ if and only if $\boldsymbol{A}$ is nonsingular
- If $\boldsymbol{A}$ is triangular (upper or lower), then $\det \boldsymbol{A} = \prod_{i=1}^{k} a_{ii}$
- If $\boldsymbol{A}$ is orthogonal, then $\det \boldsymbol{A} = \pm 1$

A.7 Eigenvalues

The characteristic equation of a $k \times k$ square matrix $\boldsymbol{A}$ is

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}_k) = 0.$$

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** or **characteristic roots** or **eigenvalues** of $\boldsymbol{A}$. If λ_i is an eigenvalue of $\boldsymbol{A}$, then $\boldsymbol{A} - \lambda_i \boldsymbol{I}_k$ is singular so there exists a non-zero vector $\boldsymbol{h}_i$ such that

$$(\boldsymbol{A} - \lambda_i \boldsymbol{I}_k)\, \boldsymbol{h}_i = \boldsymbol{0}.$$

The vector $\boldsymbol{h}_i$ is called a **latent vector** or **characteristic vector** or **eigenvector** of $\boldsymbol{A}$ corresponding to λ_i.

We now state some useful properties. Let λ_i and $\boldsymbol{h}_i$, $i = 1, ..., k$ denote the k eigenvalues and eigenvectors of a square matrix $\boldsymbol{A}$. Let $\boldsymbol{\Lambda}$ be a diagonal matrix with the characteristic roots in the diagonal, and let $\boldsymbol{H} = [\boldsymbol{h}_1 \cdots \boldsymbol{h}_k]$.

- $\det(\boldsymbol{A}) = \prod_{i=1}^{k} \lambda_i$
- $\operatorname{tr}(\boldsymbol{A}) = \sum_{i=1}^{k} \lambda_i$
- $\boldsymbol{A}$ is non-singular if and only if all its characteristic roots are non-zero.
- If $\boldsymbol{A}$ has distinct characteristic roots, there exists a nonsingular matrix $\boldsymbol{P}$ such that $\boldsymbol{A} = \boldsymbol{P}^{-1}\boldsymbol{\Lambda}\boldsymbol{P}$ and $\boldsymbol{PAP}^{-1} = \boldsymbol{\Lambda}$.
- If $\boldsymbol{A}$ is symmetric, then $\boldsymbol{A} = \boldsymbol{H\Lambda H}'$ and $\boldsymbol{H}'\boldsymbol{AH} = \boldsymbol{\Lambda}$, and the characteristic roots are all real. $\boldsymbol{A} = \boldsymbol{H\Lambda H}'$ is called the **spectral decomposition** of a matrix.
- When the eigenvalues of $k \times k$ $\boldsymbol{A}$ are real it is conventional to write them in decending order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. We also write $\lambda_{\min}(\boldsymbol{A}) = \lambda_k = \min\{\lambda_\ell\}$ and $\lambda_{\max}(\boldsymbol{A}) = \lambda_1 = \max\{\lambda_\ell\}$.

- $\lambda_{\max}(\boldsymbol{A}) = \max_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$
- $\lambda_{\min}(\boldsymbol{A}) = \min_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$
- The characteristic roots of $\boldsymbol{A}^{-1}$ are λ_1^{-1}, λ_2^{-1}, ..., λ_k^{-1}.
- The matrix $\boldsymbol{H}$ has the **orthonormal** properties $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{H}\boldsymbol{H}' = \boldsymbol{I}$.
- $\boldsymbol{H}^{-1} = \boldsymbol{H}'$ and $(\boldsymbol{H}')^{-1} = \boldsymbol{H}$
- For any $k \times 1$ vector $\boldsymbol{a}$, $\lambda_{\max}(\boldsymbol{a}\boldsymbol{a}') = \boldsymbol{a}'\boldsymbol{a}$

A.8 Positive Definiteness

We say that a $k \times k$ 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:

- 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.
- If $\boldsymbol{A}$ is positive definite, then $\boldsymbol{A}$ is non-singular and $\boldsymbol{A}^{-1}$ exists. Furthermore, $\boldsymbol{A}^{-1} > 0$.
- $\boldsymbol{A} > 0$ if and only if it is symmetric and all its characteristic roots are positive.
- By the spectral decomposition, $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ is diagonal with non-negative diagonal elements. All diagonal elements of $\boldsymbol{\Lambda}$ are strictly positive if (and only if) $\boldsymbol{A} > 0$.
- If $\boldsymbol{A} > 0$ then $\boldsymbol{A}^{-1} = \boldsymbol{H}\boldsymbol{\Lambda}^{-1}\boldsymbol{H}'$.
- If $\boldsymbol{A} \geq 0$ and $\operatorname{rank}(\boldsymbol{A}) = r < k$ then $\boldsymbol{A}^{-} = \boldsymbol{H}\boldsymbol{\Lambda}^{-}\boldsymbol{H}'$ where $\boldsymbol{A}^{-}$ is the Moore-Penrose generalized inverse, and $\boldsymbol{\Lambda}^{-} = \operatorname{diag}\left(\lambda_1^{-1}, \lambda_2^{-1}, ..., \lambda_k^{-1}, 0, ..., 0\right)$
- 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 $\boldsymbol{A}$. The matrix $\boldsymbol{B}$ need not be unique. One way to construct $\boldsymbol{B}$ is to use the spectral decomposition $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{\Lambda}$ is diagonal, and then set $\boldsymbol{B} = \boldsymbol{H}\boldsymbol{\Lambda}^{1/2}$. There is a unique square root $\boldsymbol{B}$ which is also positive semi-definite $\boldsymbol{B} \geq 0$.

A square matrix $\boldsymbol{A}$ is **idempotent** if $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. If $\boldsymbol{A}$ is idempotent and symmetric then all its characteristic roots equal either zero or one and is thus positive semi-definite. To see this, note that we can write $\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'$ where $\boldsymbol{H}$ is orthogonal and $\boldsymbol{\Lambda}$ contains the r (real) characteristic roots. Then

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A} = \boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}'\boldsymbol{H}\boldsymbol{\Lambda}\boldsymbol{H}' = \boldsymbol{H}\boldsymbol{\Lambda}^2\boldsymbol{H}'.$$

By the uniqueness of the characteristic roots, we deduce that $\boldsymbol{\Lambda}^2 = \boldsymbol{\Lambda}$ and $\lambda_i^2 = \lambda_i$ for $i = 1, ..., r$. Hence they must equal either 0 or 1. It follows that the spectral decomposition of idempotent $\boldsymbol{A}$ takes the form

$$\boldsymbol{A} = \boldsymbol{H}\begin{bmatrix} \boldsymbol{I}_{k-r} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{H}' \tag{A.5}$$

with $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_k$. Additionally, $\operatorname{tr}(\boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A})$.

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 $\boldsymbol{c}$,

$$\boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} \leq \boldsymbol{c}'\boldsymbol{c} \tag{A.6}$$

$$\boldsymbol{c}'\left(\boldsymbol{I}-\boldsymbol{A}\right)\boldsymbol{c} \leq \boldsymbol{c}'\boldsymbol{c} \tag{A.7}$$

To see this, note that

$$\boldsymbol{c}'\boldsymbol{c} = \boldsymbol{c}'\boldsymbol{A}\boldsymbol{c} + \boldsymbol{c}'\left(\boldsymbol{I}-\boldsymbol{A}\right)\boldsymbol{c}.$$

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 $\boldsymbol{c}'\left(\boldsymbol{I}-\boldsymbol{A}\right)\boldsymbol{c}$ are non-negative. Thus they must satisfy (A.6)-(A.7).

A.9 Singular Values

The singular values of a $k \times r$ real matrix $\boldsymbol{A}$ are the square roots of the eigenvalues of $\boldsymbol{A}'\boldsymbol{A}$. Thus for $j = 1, ..., r$

$$s_j = \sqrt{\lambda_j\left(\boldsymbol{A}'\boldsymbol{A}\right)}$$

Since $\boldsymbol{A}'\boldsymbol{A}$ is positive semi-definite, its eigenvalues are non-negative so the singular values are 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.10 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

$$\frac{\partial}{\partial \boldsymbol{x}} g\left(\boldsymbol{x}\right) = \left(\begin{array}{c} \frac{\partial}{\partial x_1} g\left(\boldsymbol{x}\right) \\ \vdots \\ \frac{\partial}{\partial x_k} g\left(\boldsymbol{x}\right) \end{array}\right)$$

and

$$\frac{\partial}{\partial \boldsymbol{x}'} g\left(\boldsymbol{x}\right) = \left(\begin{array}{ccc} \frac{\partial}{\partial x_1} g\left(\boldsymbol{x}\right) & \cdots & \frac{\partial}{\partial x_k} g\left(\boldsymbol{x}\right) \end{array}\right).$$

Some properties are now summarized.

- $\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{a}'\boldsymbol{x}\right) = \frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}'\boldsymbol{a}\right) = \boldsymbol{a}$
- $\frac{\partial}{\partial \boldsymbol{x}'}\left(\boldsymbol{A}\boldsymbol{x}\right) = \boldsymbol{A}$
- $\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\right) = \left(\boldsymbol{A}+\boldsymbol{A}'\right)\boldsymbol{x}$
- $\frac{\partial^2}{\partial \boldsymbol{x}\partial \boldsymbol{x}'}\left(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}\right) = \boldsymbol{A}+\boldsymbol{A}'$

A.11 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

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & & a_{1n}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2n}\boldsymbol{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\boldsymbol{B} & a_{m2}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{bmatrix}.$$

Some important properties are now summarized. These results hold for matrices for which all matrix multiplications are conformable.

- $(\boldsymbol{A} + \boldsymbol{B}) \otimes \boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{C} + \boldsymbol{B} \otimes \boldsymbol{C}$
- $(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD}$
- $\boldsymbol{A} \otimes (\boldsymbol{B} \otimes \boldsymbol{C}) = (\boldsymbol{A} \otimes \boldsymbol{B}) \otimes C$
- $(\boldsymbol{A} \otimes \boldsymbol{B})' = \boldsymbol{A}' \otimes \boldsymbol{B}'$
- $\operatorname{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) \operatorname{tr}(\boldsymbol{B})$
- If $\boldsymbol{A}$ is $m \times m$ and $\boldsymbol{B}$ is $n \times n$, $\det(\boldsymbol{A} \otimes \boldsymbol{B}) = (\det(\boldsymbol{A}))^n (\det(\boldsymbol{B}))^m$
- $(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$
- If $\boldsymbol{A} > 0$ and $\boldsymbol{B} > 0$ then $\boldsymbol{A} \otimes \boldsymbol{B} > 0$
- $\operatorname{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}' \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B})$
- $\operatorname{tr}(\boldsymbol{ABCD}) = \operatorname{vec}(\boldsymbol{D}')' (\boldsymbol{C}' \otimes \boldsymbol{A}) \operatorname{vec}(\boldsymbol{B})$

A.12 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**

$$\begin{aligned}\|\boldsymbol{a}\| &= \left(\boldsymbol{a}'\boldsymbol{a}\right)^{1/2} \\ &= \left(\sum_{i=1}^{m} a_i^2\right)^{1/2}.\end{aligned}$$

Alternative norms include the $p-$norm (for $p \geq 1$)

$$\|\boldsymbol{a}\|_p = \left(\sum_{i=1}^{m} |a_i|^p\right)^{1/p}$$

Special cases include the Euclidean norm ($p = 2$), the $1-$norm

$$\|\boldsymbol{a}\|_1 = \sum_{i=1}^{m} |a_i|$$

and the sup-norm

$$\|\boldsymbol{a}\|_\infty = \max\left(|a_1|, ..., |a_m|\right).$$

For real numbers ($m = 1$) these norms coincide.

Some standard inequalities for Euclidean space are now given. The Minkowski inequality given below establishes 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

$$g\left(\sum_{j=1}^{m} a_j x_j\right) \leq \sum_{j=1}^{m} a_j g\left(x_j\right). \tag{A.8}$$

In particular, setting $a_j = 1/m$, then

$$g\left(\frac{1}{m}\sum_{j=1}^{m} x_j\right) \leq \frac{1}{m}\sum_{j=1}^{m} g\left(x_j\right). \tag{A.9}$$

If $g(\cdot) : \mathbb{R} \to \mathbb{R}$ is concave then the inequalities in (A.8) and (A.9) 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

$$x_1^{a_1} x_2^{a_2} \cdots x_m^{a_m} \leq \sum_{j=1}^{m} a_j x_j \tag{A.10}$$

Loève's c_r Inequality. For $r > 0$,

$$\left|\sum_{j=1}^{m} a_j\right|^r \leq c_r \sum_{j=1}^{m} \left|a_j\right|^r \tag{A.11}$$

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 $\boldsymbol{b}$,

$$(\boldsymbol{a}+\boldsymbol{b})'(\boldsymbol{a}+\boldsymbol{b}) \leq 2\boldsymbol{a}'\boldsymbol{a}+2\boldsymbol{b}'\boldsymbol{b} \tag{A.12}$$

Hölder's Inequality. If $p>1$, $q>1$, and $1/p+1/q=1$, then for any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\sum_{j=1}^{m} |a_j b_j| \leq \|\boldsymbol{a}\|_p \|\boldsymbol{b}\|_q \tag{A.13}$$

Minkowski's Inequality. For any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, if $p \geq 1$, then

$$\|\boldsymbol{a}+\boldsymbol{b}\|_p \leq \|\boldsymbol{a}\|_p + \|\boldsymbol{b}\|_p \tag{A.14}$$

Schwarz Inequality. For any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$,

$$\left|\boldsymbol{a}'\boldsymbol{b}\right| \leq \|\boldsymbol{a}\| \|\boldsymbol{b}\|. \tag{A.15}$$

Proof of Jensen's Inequality (A.8). By the definition of convexity, for any $\lambda \in [0,1]$

$$g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda) g(x_2). \tag{A.16}$$

This implies

$$\begin{aligned} g\left(\sum_{j=1}^{m} a_j x_j\right) &= g\left(a_1 g(x_1) + (1-a_1)\sum_{j=2}^{m} \frac{a_j}{1-a_1} x_j\right) \\ &\leq a_1 g(x_1) + (1-a_1) g\left(\sum_{j=2}^{m} b_j x_j\right). \end{aligned}$$

where $b_j = a_j/(1-a_1)$ and $\sum_{j=2}^{m} b_j = 1$. By another application of (A.16) 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.16) we obtain (A.8). ■

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.10). ■

Proof of Loève's c_r Inequality. For $r \geq 1$ this is simply a rewriting of the finite form Jensen's inequality (A.9) 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

$$\begin{aligned}(\boldsymbol{a}+\boldsymbol{b})'(\boldsymbol{a}+\boldsymbol{b}) &= \sum_{j=1}^{m} (a_j+b_j)^2 \\ &\leq 2\sum_{j=1}^{m} a_j^2 + 2\sum_{j=1}^{m} b_j^2 \\ &= 2\boldsymbol{a}'\boldsymbol{a} + 2\boldsymbol{b}'\boldsymbol{b}\end{aligned}$$

■

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.13). ■

Proof of Minkowski's Inequality. Since $1/p + 1/q = 1$ implies $q(p-1) = p$, using the triangle inequality for real numbers and two applications of Hölder's inequality

$$\begin{aligned}\|\boldsymbol{a}+\boldsymbol{b}\|_p^p &= \sum_{j=1}^{m} |a_j+b_j|^p \\ &= \sum_{j=1}^{m} |a_j+b_j|\,|a_j+b_j|^{p-1} \\ &\leq \sum_{j=1}^{m} |a_j|\,|a_j+b_j|^{p-1} + \sum_{j=1}^{m} |b_j|\,|a_j+b_j|^{p-1} \\ &\leq \|\boldsymbol{a}\|_p \left(\sum_{j=1}^{m} |a_j+b_j|^{(p-1)q}\right)^{1/q} + \|\boldsymbol{b}\|_q \left(\sum_{j=1}^{m} |a_j+b_j|^{(p-1)q}\right)^{1/q} \\ &= \left(\|\boldsymbol{a}\|_p + \|\boldsymbol{b}\|_q\right) \|\boldsymbol{a}+\boldsymbol{b}\|_p^{p-1}\end{aligned}$$

Solving, we find (A.14). ■

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.13 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

$$\begin{aligned}\|\boldsymbol{A}\|_F &= \|\text{vec}\,(\boldsymbol{A})\| \\ &= \left(\text{tr}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} \\ &= \left(\sum_{i=1}^{m}\sum_{j=1}^{k} a_{ij}^2\right)^{1/2}.\end{aligned}$$

If an $m \times m$ matrix $\boldsymbol{A}$ is symmetric with eigenvalues λ_ℓ, $\ell = 1, ..., m$, then

$$\|\boldsymbol{A}\|_F = \left(\sum_{\ell=1}^{m} \lambda_\ell^2\right)^{1/2}.$$

To see this, by the spectral decomposition $\boldsymbol{A} = \boldsymbol{H\Lambda H}'$ with $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}$ and $\boldsymbol{\Lambda} = \text{diag}\{\lambda_1, ..., \lambda_m\}$, so

$$\|\boldsymbol{A}\|_F = \left(\text{tr}\left(\boldsymbol{H\Lambda H}'\boldsymbol{H\Lambda H}'\right)\right)^{1/2} = \left(\text{tr}\left(\boldsymbol{\Lambda\Lambda}\right)\right)^{1/2} = \left(\sum_{\ell=1}^{m} \lambda_\ell^2\right)^{1/2}. \tag{A.17}$$

A useful calculation is for any $m \times 1$ vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, using (A.1),

$$\left\|\boldsymbol{a}\boldsymbol{b}'\right\|_F = \text{tr}\left(\boldsymbol{b}\boldsymbol{a}'\boldsymbol{a}\boldsymbol{b}'\right)^{1/2} = \left(\boldsymbol{b}'\boldsymbol{b}\boldsymbol{a}'\boldsymbol{a}\right)^{1/2} = \|\boldsymbol{a}\| \, \|\boldsymbol{b}\| \tag{A.18}$$

and in particular

$$\left\|\boldsymbol{a}\boldsymbol{a}'\right\|_F = \|\boldsymbol{a}\|^2. \tag{A.19}$$

The **spectral norm** of an $m \times k$ matrix $\boldsymbol{A}$ is its largest singular value

$$\|\boldsymbol{A}\|_2 = s_{\max}\left(\boldsymbol{A}\right) = \left(\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2}$$

where $\lambda_{\max}\left(\boldsymbol{B}\right)$ 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

$$\|\boldsymbol{A}\|_2 = \max_{j \leq m} |\lambda_j|.$$

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

$$\|\boldsymbol{A}\|_2 = \left(\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} \leq \left(\sum_{j=1}^{k} \lambda_\ell\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} = \|\boldsymbol{A}\|_F$$

and

$$\|\boldsymbol{A}\|_F = \left(\sum_{j=1}^{k} \lambda_\ell\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} \leq \left(r\lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right)\right)^{1/2} = \sqrt{k}\,\|\boldsymbol{A}\|_2\,.$$

Given a vector norm $\|\boldsymbol{a}\|$ the induced matrix norm is defined as

$$\|\boldsymbol{A}\| = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}\| = \sup_{\boldsymbol{x}\neq 0} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

To see that this is a norm we need to check that it satisfies the triangle inequality. Indeed

$$\|\boldsymbol{A}+\boldsymbol{B}\| = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}+\boldsymbol{B}\boldsymbol{x}\| \leq \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}\| + \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{B}\boldsymbol{x}\| = \|\boldsymbol{A}\| + \|\boldsymbol{B}\|\,.$$

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 intereest, the matrix norm induced by the Euclidean vector norm is the spectral norm. Indeed,

$$\sup_{\boldsymbol{x}'\boldsymbol{x}=1} \|\boldsymbol{A}\boldsymbol{x}\|^2 = \sup_{\boldsymbol{x}'\boldsymbol{x}=1} \boldsymbol{x}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{x}. = \lambda_{\max}\left(\boldsymbol{A}'\boldsymbol{A}\right) = \|\boldsymbol{A}\|_S^2\,.$$

It follows that the spectral norm is consistent with the Euclidean norm, and is sub-multiplicative.

A.14 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,

$$\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\| \|\boldsymbol{B}\|\,. \tag{A.20}$$

Triangle Inequality: For any $m \times k$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, and either the Frobenius or spectral norm,

$$\|\boldsymbol{A}+\boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|\,. \tag{A.21}$$

Trace Inequality. For any $m \times m$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ such that $\boldsymbol{A}$ is symmetric and $\boldsymbol{B} \geq 0$

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) \leq \|\boldsymbol{A}\|_2 \operatorname{tr}(\boldsymbol{B}) \tag{A.22}$$

where $\|\boldsymbol{A}\|_S$ is the largest singular value of $\boldsymbol{A}$.

Quadratic Inequality. For any $m \times 1$ $\boldsymbol{b}$ and $m \times m$ symmetric matrix $\boldsymbol{A}$

$$\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b} \leq \|\boldsymbol{A}\|_2 \, \boldsymbol{b}'\boldsymbol{b} \tag{A.23}$$

Strong Schwarz Matrix Inequality. For any conformable matrices $\boldsymbol{A}$ and $\boldsymbol{B}$

$$\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_2 \, \|\boldsymbol{B}\|_F \, . \tag{A.24}$$

Norm Equivalence. For any $m \times k$ matrix $\boldsymbol{A}$ of rank r

$$\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F \leq \sqrt{r} \, \|\boldsymbol{A}\|_2 \, . \tag{A.25}$$

Eigenvalue Product Inequality. For any $m \times m$ matrices $\boldsymbol{A} \geq 0$ and $\boldsymbol{B} \geq 0$, the eigenvalues $\lambda_\ell(\boldsymbol{A}\boldsymbol{B})$ are real and satisfy

$$\lambda_{\min}(\boldsymbol{A})\,\lambda_{\min}(\boldsymbol{B}) \leq \lambda_\ell(\boldsymbol{A}\boldsymbol{B}) = \lambda_\ell\left(\boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}\right) \leq \lambda_{\max}(\boldsymbol{A})\,\lambda_{\max}(\boldsymbol{B}) \tag{A.26}$$

(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

$$\begin{aligned}
\|\boldsymbol{A}\boldsymbol{B}\|_F &= \left\| \begin{matrix} \boldsymbol{a}_1'\boldsymbol{b}_1 & \boldsymbol{a}_1'\boldsymbol{b}_2 & \cdots \\ \boldsymbol{a}_2'\boldsymbol{b}_1 & \boldsymbol{a}_2'\boldsymbol{b}_2 & \cdots \\ \vdots & \vdots & \ddots \end{matrix} \right\|_F \\
&\leq \left\| \begin{matrix} \|\boldsymbol{a}_1\|\,\|\boldsymbol{b}_1\| & \|\boldsymbol{a}_1\|\,\|\boldsymbol{b}_2\| & \cdots \\ \|\boldsymbol{a}_2\|\,\|\boldsymbol{b}_1\| & \|\boldsymbol{a}_2\|\,\|\boldsymbol{b}_2\| & \cdots \\ \vdots & \vdots & \ddots \end{matrix} \right\|_F \\
&= \left(\sum_{i=1}^m \sum_{j=1}^m \|\boldsymbol{a}_i\|^2 \|\boldsymbol{b}_j\|^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^m \|\boldsymbol{a}_i\|^2 \right)^{1/2} \left(\sum_{i=1}^m \|\boldsymbol{b}_i\|^2 \right)^{1/2} \\
&= \left(\sum_{i=1}^k \sum_{j=1}^m \boldsymbol{a}_{ji}^2 \right)^{1/2} \left(\sum_{i=1}^m \sum_{j=1}^k \|\boldsymbol{b}_{ji}\|^2 \right)^{1/2} \\
&= \|\boldsymbol{A}\|_F \, \|\boldsymbol{B}\|_F
\end{aligned}$$

■

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

$$\operatorname{tr}(\boldsymbol{A}\boldsymbol{B})=\operatorname{tr}(\boldsymbol{\Lambda}\boldsymbol{C})=\sum_{j=1}^{m}\lambda_j C_{jj}\leq\max_j|\lambda_j|\sum_{j=1}^{m}C_{jj}=\|\boldsymbol{A}\|_2\operatorname{tr}(\boldsymbol{C})$$

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 $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B})\leq\|\boldsymbol{A}\|_2\operatorname{tr}(\boldsymbol{B})$ as stated. ■

Proof of Quadratic Inequality: In the Trace Inequality set $\boldsymbol{B}=\boldsymbol{b}\boldsymbol{b}'$ and note $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B})=\boldsymbol{b}'\boldsymbol{A}\boldsymbol{b}$ 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

$$\begin{aligned}\|\boldsymbol{A}\boldsymbol{B}\|_F&=\left(\operatorname{tr}\left(\boldsymbol{B}'\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\right)\right)^{1/2}\\&=\left(\operatorname{tr}\left(\boldsymbol{A}'\boldsymbol{A}\boldsymbol{B}\boldsymbol{B}'\right)\right)^{1/2}\\&\leq\left(\left\|\boldsymbol{A}'\boldsymbol{A}\right\|_2\operatorname{tr}\left(\boldsymbol{B}\boldsymbol{B}'\right)\right)^{1/2}\\&=\|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_F.\end{aligned}$$

■

Appendix B

Probability

B.1 Foundations

The set S of all possible outcomes of an experiment is called the **sample space** for the experiment. Take the simple example of tossing a coin. There are two outcomes, heads and tails, so we can write $S = \{H, T\}$. If two coins are tossed in sequence, we can write the four outcomes as $S = \{HH, HT, TH, TT\}$.

An **event** A is any collection of possible outcomes of an experiment. An event is a subset of S, including S itself and the null set $\emptyset$. Continuing the two coin example, one event is $A = \{HH, HT\}$, the event that the first coin is heads. We say that A and B are **disjoint** or **mutually exclusive** if $A \cap B = \emptyset$. For example, the sets $\{HH, HT\}$ and $\{TH\}$ are disjoint. Furthermore, if the sets $A_1, A_2, ...$ are pairwise disjoint and $\cup_{i=1}^{\infty} A_i = S$, then the collection $A_1, A_2, ...$ is called a **partition** of S.

The following are elementary set operations:

Union: $A \cup B = \{x : x \in A \text{ or } x \in B\}$.

Intersection: $A \cap B = \{x : x \in A \text{ and } x \in B\}$.

Complement: $A^c = \{x : x \notin A\}$.

The following are useful properties of set operations.

Commutatitivity: $A \cup B = B \cup A$; $\quad A \cap B = B \cap A$.

Associativity: $A \cup (B \cup C) = (A \cup B) \cup C$; $\quad A \cap (B \cap C) = (A \cap B) \cap C$.

Distributive Laws: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $\quad A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.

DeMorgan's Laws: $(A \cup B)^c = A^c \cap B^c$; $\quad (A \cap B)^c = A^c \cup B^c$.

A **probability function** assigns probabilities (numbers between 0 and 1) to events A in S. This is straightforward when S is countable; when S is uncountable we must be somewhat more careful. A set $\mathcal{B}$ is called a **sigma algebra** (or Borel field) if $\emptyset \in \mathcal{B}$, $A \in \mathcal{B}$ implies $A^c \in \mathcal{B}$, and $A_1, A_2, ... \in \mathcal{B}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$. A simple example is $\{\emptyset, S\}$ which is known as the trivial sigma algebra. For any sample space S, let $\mathcal{B}$ be the smallest sigma algebra which contains all of the open sets in S. When S is countable, $\mathcal{B}$ is simply the collection of all subsets of S, including $\emptyset$ and S. When S is the real line, then $\mathcal{B}$ is the collection of all open and closed intervals. We call $\mathcal{B}$ the sigma algebra associated with S. We only define probabilities for events contained in $\mathcal{B}$.

We now can give the axiomatic definition of probability. Given S and $\mathcal{B}$, a probability function Pr satisfies $\Pr(S) = 1$, $\Pr(A) \geq 0$ for all $A \in \mathcal{B}$, and if $A_1, A_2, ... \in \mathcal{B}$ are pairwise disjoint, then $\Pr\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \Pr(A_i)$.

Some important properties of the probability function include the following

- $\Pr(\emptyset) = 0$
- $\Pr(A) \leq 1$
- $\Pr(A^c) = 1 - \Pr(A)$

- $\Pr(B \cap A^c) = \Pr(B) - \Pr(A \cap B)$
- $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$
- If $A \subset B$ then $\Pr(A) \leq \Pr(B)$
- Bonferroni's Inequality: $\Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1$
- Boole's Inequality: $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$

For some elementary probability models, it is useful to have simple rules to count the number of objects in a set. These counting rules are facilitated by using the binomial coefficients which are defined for nonnegative integers n and r, $n \geq r$, as

$$\binom{n}{r} = \frac{n!}{r!\,(n-r)!}.$$

When counting the number of objects in a set, there are two important distinctions. Counting may be **with replacement** or **without replacement**. Counting may be **ordered** or **unordered**. For example, consider a lottery where you pick six numbers from the set 1, 2, ..., 49. This selection is without replacement if you are not allowed to select the same number twice, and is with replacement if this is allowed. Counting is ordered or not depending on whether the sequential order of the numbers is relevant to winning the lottery. Depending on these two distinctions, we have four expressions for the number of objects (possible arrangements) of size r from n objects.

	Without Replacement	With Replacement
Ordered	$\frac{n!}{(n-r)!}$	n^r
Unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

In the lottery example, if counting is unordered and without replacement, the number of potential combinations is $\binom{49}{6} = 13,983,816$.

If $\Pr(B) > 0$ the **conditional probability** of the event A given the event B is

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}.$$

For any B, the conditional probability function is a valid probability function where S has been replaced by B. Rearranging the definition, we can write

$$\Pr(A \cap B) = \Pr(A \mid B) \Pr(B)$$

which is often quite useful. We can say that the occurrence of B has no information about the likelihood of event A when $\Pr(A \mid B) = \Pr(A)$, in which case we find

$$\Pr(A \cap B) = \Pr(A) \Pr(B) \tag{B.1}$$

We say that the events A and B are **statistically independent** when (B.1) holds. Furthermore, we say that the collection of events $A_1, ..., A_k$ are **mutually independent** when for any subset $\{A_i : i \in I\}$,

$$\Pr\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \Pr(A_i).$$

Theorem 3 *(Bayes' Rule). For any set B and any partition $A_1, A_2, ...$ of the sample space, then for each $i = 1, 2, ...$*

$$\Pr(A_i \mid B) = \frac{\Pr(B \mid A_i) \Pr(A_i)}{\sum_{j=1}^{\infty} \Pr(B \mid A_j) \Pr(A_j)}$$

B.2 Random Variables

A **random variable** X is a function from a sample space S into the real line. This induces a new sample space – the real line – and a new probability function on the real line. Typically, we denote random variables by uppercase letters such as X, and use lower case letters such as x for potential values and realized values. (This is in contrast to the notation adopted for most of the textbook.) For a random variable X we define its **cumulative distribution function** (CDF) as

$$F(x) = \Pr(X \leq x). \tag{B.2}$$

Sometimes we write this as $F_X(x)$ to denote that it is the CDF of X. A function $F(x)$ is a CDF if and only if the following three properties hold:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$
2. $F(x)$ is nondecreasing in x
3. $F(x)$ is right-continuous

We say that the random variable X is **discrete** if $F(x)$ is a step function. In the latter case, the range of X consists of a countable set of real numbers $\tau_1, ..., \tau_r$. The probability function for X takes the form

$$\Pr(X = \tau_j) = \pi_j, \qquad j = 1, ..., r \tag{B.3}$$

where $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{r} \pi_j = 1$.

We say that the random variable X is **continuous** if $F(x)$ is continuous in x. In this case $\Pr(X = \tau) = 0$ for all $\tau \in R$ so the representation (B.3) is unavailable. Instead, we represent the relative probabilities by the **probability density function** (PDF)

$$f(x) = \frac{d}{dx}F(x)$$

so that

$$F(x) = \int_{-\infty}^{x} f(u)du$$

and

$$\Pr(a \leq X \leq b) = \int_{a}^{b} f(u)du.$$

These expressions only make sense if $F(x)$ is differentiable. While there are examples of continuous random variables which do not possess a PDF, these cases are unusual and are typically ignored.

A function $f(x)$ is a PDF if and only if $f(x) \geq 0$ for all $x \in R$ and $\int_{-\infty}^{\infty} f(x)dx = 1$.

B.3 Expectation

For any measurable real function g, we define the **mean** or **expectation** $\mathbb{E}g(X)$ as follows. If X is discrete,

$$\mathbb{E}g(X) = \sum_{j=1}^{r} g(\tau_j)\pi_j,$$

and if X is continuous

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

The latter is well defined and finite if

$$\int_{-\infty}^{\infty} |g(x)| f(x)dx < \infty. \tag{B.4}$$

If (B.4) does not hold, evaluate

$$I_1 = \int_{g(x)>0} g(x)f(x)dx$$
$$I_2 = -\int_{g(x)<0} g(x)f(x)dx$$

If $I_1 = \infty$ and $I_2 < \infty$ then we define $\mathbb{E}g(X) = \infty$. If $I_1 < \infty$ and $I_2 = \infty$ then we define $\mathbb{E}g(X) = -\infty$. If both $I_1 = \infty$ and $I_2 = \infty$ then $\mathbb{E}g(X)$ is undefined.

Since $\mathbb{E}\left(a + bX\right) = a + b\mathbb{E}X$, we say that expectation is a linear operator.

For $m > 0$, we define the $m'th$ **moment** of X as $\mathbb{E}X^m$ and the $m'th$ **central moment** as $\mathbb{E}\left(X - \mathbb{E}X\right)^m$.

Two special moments are the **mean** $\mu = \mathbb{E}X$ and **variance** $\sigma^2 = \mathbb{E}\left(X-\mu\right)^2 = \mathbb{E}X^2 - \mu^2$. We call $\sigma = \sqrt{\sigma^2}$ the **standard deviation** of X. We can also write $\sigma^2 = \text{var}(X)$. For example, this allows the convenient expression $\text{var}(a + bX) = b^2\,\text{var}(X)$.

The **moment generating function** (MGF) of X is

$$M(\lambda) = \mathbb{E}\exp\left(\lambda X\right).$$

The MGF does not necessarily exist. However, when it does and $\mathbb{E}\left|X\right|^m < \infty$ then

$$\left.\frac{d^m}{d\lambda^m}M(\lambda)\right|_{\lambda=0} = \mathbb{E}\left(X^m\right)$$

which is why it is called the moment generating function.

More generally, the **characteristic function** (CF) of X is

$$C(\lambda) = \mathbb{E}\exp\left(\mathrm{i}\lambda X\right)$$

where $\mathrm{i} = \sqrt{-1}$ is the imaginary unit. The CF always exists, and when $\mathbb{E}\left|X\right|^m < \infty$

$$\left.\frac{d^m}{d\lambda^m}C(\lambda)\right|_{\lambda=0} = \mathrm{i}^m\mathbb{E}\left(X^m\right).$$

The L^p **norm**, $p \geq 1$, of the random variable X is

$$\left\|X\right\|_p = \left(\mathbb{E}\left|X\right|^p\right)^{1/p}.$$

B.4 Gamma Function

The gamma function is defined for $\alpha > 0$ as

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\exp\left(-x\right).$$

It satisfies the property

$$\Gamma(1+\alpha) = \Gamma(\alpha)\alpha$$

so for positive integers n,

$$\Gamma(n) = (n-1)!$$

Special values include

$$\Gamma\left(1\right) = 1$$

and

$$\Gamma\left(\frac{1}{2}\right) = \pi^{1/2}.$$

Sterling's formula is an expansion for the its logarithm

$$\log\Gamma(\alpha) = \frac{1}{2}\log(2\pi) + \left(\alpha - \frac{1}{2}\right)\log\alpha - z + \frac{1}{12\alpha} - \frac{1}{360\alpha^3} + \frac{1}{1260\alpha^5} + \cdots$$

B.5 Common Distributions

For reference, we now list some important discrete distribution function.
Bernoulli

$$\begin{aligned}\Pr\left(X=x\right) &= p^x(1-p)^{1-x}, \qquad x=0,1; \qquad 0\le p\le 1\\ \mathbb{E}X &= p\\ \operatorname{var}(X) &= p(1-p)\end{aligned}$$

Binomial

$$\begin{aligned}\Pr\left(X=x\right) &= \binom{n}{x}p^x\left(1-p\right)^{n-x}, \qquad x=0,1,...,n; \qquad 0\le p\le 1\\ \mathbb{E}X &= np\\ \operatorname{var}(X) &= np(1-p)\end{aligned}$$

Geometric

$$\begin{aligned}\Pr\left(X=x\right) &= p(1-p)^{x-1}, \qquad x=1,2,...; \qquad 0\le p\le 1\\ \mathbb{E}X &= \frac{1}{p}\\ \operatorname{var}(X) &= \frac{1-p}{p^2}\end{aligned}$$

Multinomial

$$\begin{aligned}\Pr\left(X_1=x_1,X_2=x_2,...,X_m=x_m\right) &= \frac{n!}{x_1!x_2!\cdots x_m!}p_1^{x_1}p_2^{x_2}\cdots p_m^{x_m},\\ x_1+\cdots+x_m &= n;\\ p_1+\cdots+p_m &= 1\\ \mathbb{E}X_i &= p_i\\ \operatorname{var}(X_i) &= np_i(1-p_i)\\ \operatorname{cov}\left(X_i,X_j\right) &= -np_ip_j\end{aligned}$$

Negative Binomial

$$\begin{aligned}\Pr\left(X=x\right) &= \frac{\Gamma\left(r+x\right)}{x!\Gamma\left(r\right)}p^r(1-p)^{x-1}, \qquad x=0,1,2,...; \qquad 0\le p\le 1\\ \mathbb{E}X&=\frac{r\left(1-p\right)}{p}\\ \operatorname{var}(X) &= \frac{r\left(1-p\right)}{p^2}\end{aligned}$$

Poisson

$$\begin{aligned}\Pr\left(X=x\right) &= \frac{\exp\left(-\lambda\right)\lambda^x}{x!}, \qquad x=0,1,2,..., \qquad \lambda>0\\ \mathbb{E}X &= \lambda\\ \operatorname{var}(X) &= \lambda\end{aligned}$$

We now list some important continuous distributions.

Beta

$$\begin{aligned} f(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 \le x \le 1; \qquad \alpha > 0,\ \beta > 0 \\ \mu &= \frac{\alpha}{\alpha+\beta} \\ \operatorname{var}(X) &= \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2} \end{aligned}$$

Cauchy

$$\begin{aligned} f(x) &= \frac{1}{\pi\left(1+x^2\right)}, \qquad -\infty < x < \infty \\ \mathbb{E}X &= \infty \\ \operatorname{var}(X) &= \infty \end{aligned}$$

Exponential

$$\begin{aligned} f(x) &= \frac{1}{\theta}\exp\left(\frac{x}{\theta}\right), \qquad 0 \le x < \infty; \qquad \theta > 0 \\ \mathbb{E}X &= \theta \\ \operatorname{var}(X) &= \theta^2 \end{aligned}$$

Logistic

$$\begin{aligned} f(x) &= \frac{\exp(-x)}{(1+\exp(-x))^2}, \qquad -\infty < x < \infty; \\ \mathbb{E}X &= 0 \\ \operatorname{var}(X) &= \frac{\pi^2}{3} \end{aligned}$$

Lognormal

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma x}\exp\left(-\frac{(\log x-\mu)^2}{2\sigma^2}\right), \qquad 0 \le x < \infty; \qquad \sigma > 0 \\ \mathbb{E}X &= \exp\left(\mu+\sigma^2/2\right) \\ \operatorname{var}(X) &= \exp\left(2\mu+2\sigma^2\right)-\exp\left(2\mu+\sigma^2\right) \end{aligned}$$

Pareto

$$\begin{aligned} f(x) &= \frac{\beta\alpha^\beta}{x^{\beta+1}}, \qquad \alpha \le x < \infty, \qquad \alpha > 0, \qquad \beta > 0 \\ \mathbb{E}X &= \frac{\beta\alpha}{\beta-1}, \qquad \beta > 1 \\ \operatorname{var}(X) &= \frac{\beta\alpha^2}{(\beta-1)^2(\beta-2)}, \qquad \beta > 2 \end{aligned}$$

Uniform

$$\begin{aligned} f(x) &= \frac{1}{b-a}, \qquad a \le x \le b \\ \mathbb{E}X &= \frac{a+b}{2} \\ \operatorname{var}(X) &= \frac{(b-a)^2}{12} \end{aligned}$$

Weibull

$$\begin{aligned} f(x) &= \frac{\gamma}{\beta} x^{\gamma-1} \exp\left(-\frac{x^\gamma}{\beta}\right), \qquad 0 \le x < \infty; \qquad \gamma > 0,\ \beta > 0 \\ \mathbb{E}X &= \beta^{1/\gamma}\Gamma\left(1+\frac{1}{\gamma}\right) \\ \text{var}(X) &= \beta^{2/\gamma}\left(\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma^2\left(1+\frac{1}{\gamma}\right)\right) \end{aligned}$$

Gamma

$$\begin{aligned} f(x) &= \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} \exp\left(-\frac{x}{\theta}\right), \qquad 0 \le x < \infty; \qquad \alpha > 0,\ \theta > 0 \\ \mathbb{E}X &= \alpha\theta \\ \text{var}(X) &= \alpha\theta^2 \end{aligned}$$

Chi-Square

$$\begin{aligned} f(x) &= \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} \exp\left(-\frac{x}{2}\right), \qquad 0 \le x < \infty; \qquad r > 0 \\ \mathbb{E}X &= r \\ \text{var}(X) &= 2r \end{aligned}$$

Normal

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty; \qquad -\infty < \mu < \infty,\ \sigma^2 > 0 \\ \mathbb{E}X &= \mu \\ \text{var}(X) &= \sigma^2 \end{aligned}$$

Student t

$$\begin{aligned} f(x) &= \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)} \left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)}, \qquad -\infty < x < \infty; \qquad r > 0 \\ \mathbb{E}X &= 0 \text{ if } r > 1 \\ \text{var}(X) &= \frac{r}{r-2} \text{ if } r > 2 \end{aligned}$$

B.6 Multivariate Random Variables

A pair of bivariate random variables (X, Y) is a function from the sample space into $\mathbb{R}^2$. The joint CDF of (X, Y) is

$$F(x, y) = \Pr(X \le x, Y \le y).$$

If F is continuous, the joint probability density function is

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y).$$

For a Borel measurable set $A \in R^2$,

$$\Pr((X, Y) \in A) = \int\int_A f(x, y) dx dy$$

For any measurable function $g(x, y)$,

$$\mathbb{E}g(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

The **marginal distribution** of X is

$$\begin{aligned} F_X(x) &= \Pr(X \le x) \\ &= \lim_{y \to \infty} F(x, y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x, y) dy dx \end{aligned}$$

so the **marginal density** of X is

$$f_X(x) = \frac{d}{dx} F_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

Similarly, the marginal density of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

The random variables X and Y are defined to be **independent** if $f(x, y) = f_X(x) f_Y(y)$. Furthermore, X and Y are independent if and only if there exist functions $g(x)$ and $h(y)$ such that $f(x, y) = g(x)h(y)$.

If X and Y are independent, then

$$\begin{aligned} \mathbb{E}\left(g(X)h(Y)\right) &= \int \int g(x)h(y) f(y, x) dy dx \\ &= \int \int g(x)h(y) f_Y(y) f_X(x) dy dx \\ &= \int g(x) f_X(x) dx \int h(y) f_Y(y) dy \\ &= \mathbb{E}g\left(X\right) \mathbb{E}h\left(Y\right). \end{aligned} \tag{B.5}$$

if the expectations exist. For example, if X and Y are independent then

$$\mathbb{E}(XY) = \mathbb{E}X\mathbb{E}Y.$$

Another implication of (B.5) is that if X and Y are independent and $Z = X + Y$, then

$$\begin{aligned} M_Z(\lambda) &= \mathbb{E}\exp\left(\lambda\left(X + Y\right)\right) \\ &= \mathbb{E}\left(\exp\left(\lambda X\right)\exp\left(\lambda Y\right)\right) \\ &= \mathbb{E}\exp\left(\lambda' X\right) \mathbb{E}\exp\left(\lambda' Y\right) \\ &= M_X(\lambda) M_Y(\lambda). \end{aligned} \tag{B.6}$$

The covariance between X and Y is

$$\operatorname{cov}(X, Y) = \sigma_{XY} = \mathbb{E}\left(\left(X - \mathbb{E}X\right)\left(Y - \mathbb{E}Y\right)\right) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y.$$

The correlation between X and Y is

$$\operatorname{corr}\left(X, Y\right) = \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

The Cauchy-Schwarz Inequality implies that

$$|\rho_{XY}| \leq 1. \tag{B.7}$$

The correlation is a measure of linear dependence, free of units of measurement.

If X and Y are independent, then $\sigma_{XY} = 0$ and $\rho_{XY} = 0$. The reverse, however, is not true. For example, if $\mathbb{E}X = 0$ and $\mathbb{E}X^3 = 0$, then $\operatorname{cov}(X, X^2) = 0$.

A useful fact is that

$$\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X,Y).$$

An implication is that if X and Y are independent, then

$$\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y),$$

the variance of the sum is the sum of the variances.

A $k \times 1$ random vector $\boldsymbol{X} = (X_1, ..., X_k)'$ is a function from S to $\mathbb{R}^k$. Let $\boldsymbol{x} = (x_1, ..., x_k)'$ denote a vector in $\mathbb{R}^k$. (In this Appendix, we use bold to denote vectors. Bold capitals $\boldsymbol{X}$ are random vectors and bold lower case $\boldsymbol{x}$ are nonrandom vectors. Again, this is in distinction to the notation used in the bulk of the text) The vector $\boldsymbol{X}$ has the distribution and density functions

$$\begin{aligned} F(\boldsymbol{x}) &= \Pr(\boldsymbol{X} \leq \boldsymbol{x}) \\ f(\boldsymbol{x}) &= \frac{\partial^k}{\partial x_1 \cdots \partial x_k} F(\boldsymbol{x}). \end{aligned}$$

For a measurable function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^s$, we define the expectation

$$\mathbb{E}\boldsymbol{g}(\boldsymbol{X}) = \int_{\mathbb{R}^k} g(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x}$$

where the symbol $d\boldsymbol{x}$ denotes $dx_1 \cdots dx_k$. In particular, we have the $k \times 1$ multivariate mean

$$\boldsymbol{\mu} = \mathbb{E}\boldsymbol{X}$$

and $k \times k$ covariance matrix

$$\begin{aligned} \boldsymbol{\Sigma} &= \mathbb{E}\left((\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})'\right) \\ &= \mathbb{E}\boldsymbol{X}\boldsymbol{X}' - \boldsymbol{\mu}\boldsymbol{\mu}' \end{aligned}$$

If the elements of $\boldsymbol{X}$ are mutually independent, then $\boldsymbol{\Sigma}$ is a diagonal matrix and

$$\operatorname{var}\left(\sum_{i=1}^k \boldsymbol{X}_i\right) = \sum_{i=1}^k \operatorname{var}(\boldsymbol{X}_i)$$

B.7 Conditional Distributions and Expectation

The **conditional density** of Y given $\boldsymbol{X} = \boldsymbol{x}$ is defined as

$$f_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x}, y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

if $f_{\boldsymbol{X}}(\boldsymbol{x})>0$. One way to derive this expression from the definition of conditional probability is

$$\begin{aligned}
f_{Y|X}(y\mid\boldsymbol{x}) &= \frac{\partial}{\partial y}\lim_{\varepsilon\to 0}\Pr\left(Y\le y\mid \boldsymbol{x}\le\boldsymbol{X}\le\boldsymbol{x}+\varepsilon\right)\\
&= \frac{\partial}{\partial y}\lim_{\varepsilon\to 0}\frac{\Pr\left(\{Y\le y\}\cap\{\boldsymbol{x}\le\boldsymbol{X}\le\boldsymbol{x}+\varepsilon\}\right)}{\Pr(\boldsymbol{x}\le X\le\boldsymbol{x}+\varepsilon)}\\
&= \frac{\partial}{\partial y}\lim_{\varepsilon\to 0}\frac{F(\boldsymbol{x}+\varepsilon,y)-F(\boldsymbol{x},y)}{F_{\boldsymbol{X}}(\boldsymbol{x}+\varepsilon)-F_X(\boldsymbol{x})}\\
&= \frac{\partial}{\partial y}\lim_{\varepsilon\to 0}\frac{\frac{\partial}{\partial x}F(\boldsymbol{x}+\varepsilon,y)}{f_{\boldsymbol{X}}(\boldsymbol{x}+\varepsilon)}\\
&= \frac{\frac{\partial^2}{\partial x\partial y}F(\boldsymbol{x},y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}\\
&= \frac{f(\boldsymbol{x},y)}{f_{\boldsymbol{X}}(\boldsymbol{x})}.
\end{aligned}$$

The **conditional mean** or **conditional expectation** is the function

$$m(\boldsymbol{x})=\mathbb{E}\left(Y\mid\boldsymbol{X}=\boldsymbol{x}\right)=\int_{-\infty}^{\infty}yf_{Y|\boldsymbol{X}}\left(y\mid\boldsymbol{x}\right)dy.$$

The conditional mean $m(\boldsymbol{x})$ is a function, meaning that when $\boldsymbol{X}$ equals $\boldsymbol{x}$, then the expected value of Y is $m(\boldsymbol{x})$.

Similarly, we define the conditional variance of Y given $\boldsymbol{X}=\boldsymbol{x}$ as

$$\begin{aligned}
\sigma^2(\boldsymbol{x}) &= \operatorname{var}\left(Y\mid\boldsymbol{X}=\boldsymbol{x}\right)\\
&= \mathbb{E}\left((Y-m(\boldsymbol{x}))^2\mid\boldsymbol{X}=\boldsymbol{x}\right)\\
&= \mathbb{E}\left(Y^2\mid X=\boldsymbol{x}\right)-m(\boldsymbol{x})^2.
\end{aligned}$$

Evaluated at $\boldsymbol{x}=\boldsymbol{X}$, the conditional mean $m(\boldsymbol{X})$ and conditional variance $\sigma^2(\boldsymbol{X})$ are random variables, functions of $\boldsymbol{X}$. We write this as $\mathbb{E}(Y\mid\boldsymbol{X})=m(\boldsymbol{X})$ and $\operatorname{var}(Y\mid\boldsymbol{X})=\sigma^2(\boldsymbol{X})$. For example, if $\mathbb{E}\left(Y\mid\boldsymbol{X}=\boldsymbol{x}\right)=\alpha+\beta'\boldsymbol{x}$, then $\mathbb{E}\left(Y\mid\boldsymbol{X}\right)=\alpha+\beta'\boldsymbol{X}$, a transformation of $\boldsymbol{X}$.

The following are important facts about conditional expectations.

Simple Law of Iterated Expectations:

$$\mathbb{E}\left(\mathbb{E}\left(Y\mid\boldsymbol{X}\right)\right)=\mathbb{E}\left(Y\right)\tag{B.8}$$

Proof:

$$\begin{aligned}
\mathbb{E}\left(\mathbb{E}\left(Y\mid\boldsymbol{X}\right)\right) &= \mathbb{E}\left(m(\boldsymbol{X})\right)\\
&= \int_{-\infty}^{\infty}m(\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf_{Y|\boldsymbol{X}}\left(y\mid\boldsymbol{x}\right)f_{\boldsymbol{X}}(\boldsymbol{x})dyd\boldsymbol{x}\\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}yf\left(y,\boldsymbol{x}\right)dyd\boldsymbol{x}\\
&= \mathbb{E}(Y).
\end{aligned}$$

Law of Iterated Expectations:

$$\mathbb{E}\left(\mathbb{E}\left(Y\mid\boldsymbol{X},\boldsymbol{Z}\right)\mid\boldsymbol{X}\right)=\mathbb{E}\left(Y\mid\boldsymbol{X}\right)\tag{B.9}$$

Conditioning Theorem. For any function $g(\boldsymbol{x})$,

$$\mathbb{E}\left(g(\boldsymbol{X})Y \mid \boldsymbol{X}\right) = g\left(\boldsymbol{X}\right)\mathbb{E}\left(Y \mid \boldsymbol{X}\right) \tag{B.10}$$

Proof: Let

$$\begin{aligned} h(\boldsymbol{x}) &= \mathbb{E}\left(g(\boldsymbol{X})Y \mid \boldsymbol{X} = \boldsymbol{x}\right) \\ &= \int_{-\infty}^{\infty} g(\boldsymbol{x}) y f_{Y|\boldsymbol{X}}\left(y \mid \boldsymbol{x}\right) dy \\ &= g(\boldsymbol{x}) \int_{-\infty}^{\infty} y f_{Y|\boldsymbol{X}}\left(y \mid \boldsymbol{x}\right) dy \\ &= g(\boldsymbol{x}) m(\boldsymbol{x}) \end{aligned}$$

where $m(\boldsymbol{x}) = \mathbb{E}\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right)$. Thus $h(\boldsymbol{X}) = g(\boldsymbol{X})m(\boldsymbol{X})$, which is the same as $\mathbb{E}\left(g(\boldsymbol{X})Y \mid \boldsymbol{X}\right) = g\left(\boldsymbol{X}\right)\mathbb{E}\left(Y \mid \boldsymbol{X}\right)$.

B.8 Transformations

Suppose that $\boldsymbol{X} \in \mathbb{R}^k$ with continuous distribution function $F_{\boldsymbol{X}}(\boldsymbol{x})$ and density $f_{\boldsymbol{X}}(\boldsymbol{x})$. Let $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X})$ where $\boldsymbol{g}(\boldsymbol{x}) : \mathbb{R}^k \to \mathbb{R}^k$ is one-to-one, differentiable, and invertible. Let $\boldsymbol{h}(\boldsymbol{y})$ denote the inverse of $\boldsymbol{g}(\boldsymbol{x})$. The **Jacobian** is

$$J(\boldsymbol{y}) = \det\left(\frac{\partial}{\partial \boldsymbol{y}'}\boldsymbol{h}(\boldsymbol{y})\right).$$

Consider the univariate case $k = 1$. If $g(x)$ is an increasing function, then $g(X) \leq Y$ if and only if $X \leq h(Y)$, so the distribution function of Y is

$$\begin{aligned} F_Y(y) &= \Pr\left(g(X) \leq y\right) \\ &= \Pr\left(X \leq h(Y)\right) \\ &= F_X\left(h(Y)\right). \end{aligned}$$

Taking the derivative, the density of Y is

$$f_Y(y) = \frac{d}{dy}F_Y(y) = f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

If $g(x)$ is a decreasing function, then $g(X) \leq Y$ if and only if $X \geq h(Y)$, so

$$\begin{aligned} F_Y(y) &= \Pr\left(g(X) \leq y\right) \\ &= 1 - \Pr\left(X \geq h(Y)\right) \\ &= 1 - F_X\left(h(Y)\right) \end{aligned}$$

and the density of Y is

$$f_Y(y) = -f_X\left(h(Y)\right)\frac{d}{dy}h(y).$$

We can write these two cases jointly as

$$f_Y(y) = f_X\left(h(Y)\right)\left|J(y)\right|. \tag{B.11}$$

This is known as the **change-of-variables** formula. This same formula (B.11) holds for $k > 1$, but its justification requires deeper results from analysis.

As one example, take the case $X \sim U[0,1]$ and $Y = -\log(X)$. Here, $g(x) = -\log(x)$ and $h(y) = \exp(-y)$ so the Jacobian is $J(y) = -\exp(y)$. As the range of X is $[0,1]$, that for Y is $[0,\infty)$. Since $f_X\left(x\right) = 1$ for $0 \leq x \leq 1$ (B.11) shows that

$$f_Y(y) = \exp(-y), \qquad 0 \leq y \leq \infty,$$

an exponential density.

B.9 Normal and Related Distributions

The **standard normal** density is

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad -\infty < x < \infty.$$

It is conventional to write $X \sim \mathrm{N}(0,1)$, and to denote the standard normal density function by $\phi(x)$ and its distribution function by $\Phi(x)$. The latter has no closed-form solution. The normal density has all moments finite. Since it is symmetric about zero all odd moments are zero. By iterated integration by parts, we can also show that $\mathbb{E}X^2 = 1$ and $\mathbb{E}X^4 = 3$. In fact, for any positive integer m, $\mathbb{E}X^{2m} = (2m-1)!! = (2m-1)\cdot(2m-3)\cdots 1$. Thus $\mathbb{E}X^4 = 3$, $\mathbb{E}X^6 = 15$, $\mathbb{E}X^8 = 105$, and $\mathbb{E}X^{10} = 945$.

If Z is standard normal and $X = \mu + \sigma Z$, then using the change-of-variables formula, X has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

which is the **univariate normal density**. The mean and variance of the distribution are μ and σ^2, and it is conventional to write $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$.

For $\boldsymbol{x} \in \mathbb{R}^k$, the **multivariate normal density** is

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right), \qquad \boldsymbol{x} \in \mathbb{R}^k.$$

The mean and covariance matrix of the distribution are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, and it is conventional to write $X \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The MGF and CF of the multivariate normal are $\exp\left(\boldsymbol{\lambda}'\boldsymbol{\mu} + \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}/2\right)$ and $\exp\left(\mathrm{i}\boldsymbol{\lambda}'\boldsymbol{\mu} - \boldsymbol{\lambda}'\boldsymbol{\Sigma}\boldsymbol{\lambda}/2\right)$, respectively.

If $\boldsymbol{X} \in \mathbb{R}^k$ is multivariate normal and the elements of $\boldsymbol{X}$ are mutually uncorrelated, then $\boldsymbol{\Sigma} = \mathrm{diag}\{\sigma_j^2\}$ is a diagonal matrix. In this case the density function can be written as

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{(2\pi)^{k/2}\sigma_1\cdots\sigma_k} \exp\left(-\left(\frac{(x_1-\mu_1)^2/\sigma_1^2 + \cdots + (x_k-\mu_k)^2/\sigma_k^2}{2}\right)\right) \\ &= \prod_{j=1}^{k} \frac{1}{(2\pi)^{1/2}\sigma_j} \exp\left(-\frac{(x_j-\mu_j)^2}{2\sigma_j^2}\right) \end{aligned}$$

which is the product of marginal univariate normal densities. This shows that if $\boldsymbol{X}$ is multivariate normal with uncorrelated elements, then they are mutually independent.

Theorem B.9.1 *If $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$ with $\boldsymbol{B}$ an invertible matrix, then $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}, \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')$.*

Theorem B.9.2 *Let $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_r)$. Then $Q = \boldsymbol{X}'\boldsymbol{X}$ is distributed chi-square with r degrees of freedom, written χ_r^2.*

Theorem B.9.3 *If $\boldsymbol{Z} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{A})$ with $\boldsymbol{A} > 0$, $q \times q$, then $\boldsymbol{Z}'\boldsymbol{A}^{-1}\boldsymbol{Z} \sim \chi_q^2$.*

Theorem B.9.4 *Let $Z \sim \mathrm{N}(0,1)$ and $Q \sim \chi_r^2$ be independent. Then $T_r = Z/\sqrt{Q/r}$ is distributed as student's t with r degrees of freedom.*

Proof of Theorem B.9.1. By the change-of-variables formula, the density of $\boldsymbol{Y} = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}$ is

$$f(\boldsymbol{y}) = \frac{1}{(2\pi)^{k/2} \det(\boldsymbol{\Sigma}_Y)^{1/2}} \exp\left(-\frac{(\boldsymbol{y} - \boldsymbol{\mu}_Y)' \boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_Y)}{2} \right), \qquad \boldsymbol{y} \in \mathbb{R}^k.$$

where $\boldsymbol{\mu}_Y = \boldsymbol{a} + \boldsymbol{B}\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_Y = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}'$, where we used the fact that $\det(\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}')^{1/2} = \det(\boldsymbol{\Sigma})^{1/2} \det(\boldsymbol{B})$.
■

Proof of Theorem B.9.2. First, suppose a random variable Q is distributed chi-square with r degrees of freedom. It has the MGF

$$\mathbb{E} \exp(tQ) = \int_0^\infty \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{r/2}} x^{r/2-1} \exp(tx) \exp(-x/2)\, dy = (1-2t)^{-r/2}$$

where the second equality uses the fact that $\int_0^\infty y^{a-1} \exp(-by)\, dy = b^{-a}\Gamma(a)$, which can be found by applying change-of-variables to the gamma function. Our goal is to calculate the MGF of $Q = \boldsymbol{X}'\boldsymbol{X}$ and show that it equals $(1-2t)^{-r/2}$, which will establish that $Q \sim \chi_r^2$.

Note that we can write $Q = \boldsymbol{X}'\boldsymbol{X} = \sum_{j=1}^r Z_j^2$ where the Z_j are independent $\mathrm{N}(0,1)$. The distribution of each of the Z_j^2 is

$$\begin{aligned}
\Pr\left(Z_j^2 \le y\right) &= 2\Pr\left(0 \le Z_j \le \sqrt{y}\right) \\
&= 2\int_0^{\sqrt{y}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\
&= \int_0^y \frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{1/2}} s^{-1/2} \exp\left(-\frac{s}{2}\right) ds
\end{aligned}$$

using the change–of-variables $s = x^2$ and the fact $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. Thus the density of Z_j^2 is

$$f_1(x) = \frac{1}{\Gamma\left(\frac{1}{2}\right) 2^{1/2}} x^{-1/2} \exp\left(-\frac{x}{2}\right)$$

which is the χ_1^2 and by our above calculation has the MGF of $\mathbb{E}\exp\left(tZ_j^2\right) = (1-2t)^{-1/2}$.

Since the Z_j^2 are mutually independent, (B.6) implies that the MGF of $Q = \sum_{j=1}^r Z_j^2$ is $\left[(1-2t)^{-1/2}\right]^r = (1-2t)^{-r/2}$, which is the MGF of the χ_r^2 density as desired. ■

Proof of Theorem B.9.3. The fact that $\boldsymbol{A} > 0$ means that we can write $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}'$ where $\boldsymbol{C}$ is non-singular. Then $\boldsymbol{A}^{-1} = \boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}$ and

$$\boldsymbol{C}^{-1}\boldsymbol{Z} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{C}^{-1}\boldsymbol{A}\boldsymbol{C}^{-1\prime}\right) = \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{C}^{-1}\boldsymbol{C}\boldsymbol{C}'\boldsymbol{C}^{-1\prime}\right) = \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_q\right).$$

Thus

$$\boldsymbol{Z}'\boldsymbol{A}^{-1}\boldsymbol{Z} = \boldsymbol{Z}'\boldsymbol{C}^{-1\prime}\boldsymbol{C}^{-1}\boldsymbol{Z} = \left(\boldsymbol{C}^{-1}\boldsymbol{Z}\right)'\left(\boldsymbol{C}^{-1}\boldsymbol{Z}\right) \sim \chi_q^2.$$

■

Proof of Theorem B.9.4. Using the simple law of iterated expectations, T_r has distribution function

$$\begin{aligned} F(x) &= \Pr\left(\frac{Z}{\sqrt{Q/r}} \leq x\right) \\ &= \mathbb{E}\left\{Z \leq x\sqrt{\frac{Q}{r}}\right\} \\ &= \mathbb{E}\left[\Pr\left(Z \leq x\sqrt{\frac{Q}{r}} \mid Q\right)\right] \\ &= \mathbb{E}\Phi\left(x\sqrt{\frac{Q}{r}}\right) \end{aligned}$$

Thus its density is

$$\begin{aligned} f(x) &= \mathbb{E}\frac{d}{dx}\Phi\left(x\sqrt{\frac{Q}{r}}\right) \\ &= \mathbb{E}\left(\phi\left(x\sqrt{\frac{Q}{r}}\right)\sqrt{\frac{Q}{r}}\right) \\ &= \int_0^\infty \left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{qx^2}{2r}\right)\right)\sqrt{\frac{q}{r}}\left(\frac{1}{\Gamma\left(\frac{r}{2}\right)2^{r/2}}q^{r/2-1}\exp\left(-q/2\right)\right)dq \\ &= \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{r\pi}\Gamma\left(\frac{r}{2}\right)}\left(1+\frac{x^2}{r}\right)^{-\left(\frac{r+1}{2}\right)} \end{aligned}$$

which is that of the student t with r degrees of freedom. ■

B.10 Inequalities

Jensen's Inequality. If $g(\cdot) : \mathbb{R}^m \to \mathbb{R}$ is convex, then for any random vector $\boldsymbol{x}$ for which $\mathbb{E}\left\|\boldsymbol{x}\right\| < \infty$ and $\mathbb{E}\left|g(\boldsymbol{x})\right| < \infty$,

$$g(\mathbb{E}(\boldsymbol{x})) \leq \mathbb{E}\left(g(\boldsymbol{x})\right). \tag{B.12}$$

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 $\mathbb{E}\left\|\boldsymbol{y}\right\| < \infty$ and $\mathbb{E}\left\|g(\boldsymbol{y})\right\| < \infty$,

$$g(\mathbb{E}(\boldsymbol{y} \mid \boldsymbol{x})) \leq \mathbb{E}\left(g(\boldsymbol{y}) \mid \boldsymbol{x}\right). \tag{B.13}$$

Conditional Expectation Inequality. For any $r \geq 1$ such that $\mathbb{E}\left|y\right|^r < \infty$, then

$$\mathbb{E}\left|\mathbb{E}(y \mid \boldsymbol{x})\right|^r \leq \mathbb{E}\left|y\right|^r < \infty. \tag{B.14}$$

Expectation Inequality. For any random matrix $\boldsymbol{Y}$ for which $\mathbb{E}\left\|\boldsymbol{Y}\right\| < \infty$,

$$\left\|\mathbb{E}(\boldsymbol{Y})\right\| \leq \mathbb{E}\left\|\boldsymbol{Y}\right\|. \tag{B.15}$$

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 $\boldsymbol{Y}$,

$$\mathbb{E}\left\|\boldsymbol{X}'\boldsymbol{Y}\right\| \leq \left(\mathbb{E}\left\|\boldsymbol{X}\right\|^p\right)^{1/p}\left(\mathbb{E}\left\|\boldsymbol{Y}\right\|^q\right)^{1/q}. \tag{B.16}$$

Cauchy-Schwarz Inequality. For any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\mathbb{E}\left\|\boldsymbol{X}'\boldsymbol{Y}\right\| \leq \left(\mathbb{E}\left\|\boldsymbol{X}\right\|^2\right)^{1/2}\left(\mathbb{E}\left\|\boldsymbol{Y}\right\|^2\right)^{1/2}. \tag{B.17}$$

Matrix Cauchy-Schwarz Inequality. Tripathi (1999). For any random $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{y} \in \mathbb{R}^\ell$,

$$\mathbb{E}\boldsymbol{y}\boldsymbol{x}'\left(\mathbb{E}\boldsymbol{x}\boldsymbol{x}'\right)^{-}\mathbb{E}\boldsymbol{x}\boldsymbol{y}' \leq \mathbb{E}\boldsymbol{y}\boldsymbol{y}' \tag{B.18}$$

Minkowski's Inequality. For any random $m \times n$ matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$,

$$\left(\mathbb{E}\left\|\boldsymbol{X}+\boldsymbol{Y}\right\|^p\right)^{1/p} \leq \left(\mathbb{E}\left\|\boldsymbol{X}\right\|^p\right)^{1/p} + \left(\mathbb{E}\left\|\boldsymbol{Y}\right\|^p\right)^{1/p} \tag{B.19}$$

Liapunov's Inequality. For any random $m \times n$ matrix $\boldsymbol{X}$ and $1 \leq r \leq p$,

$$\left(\mathbb{E}\left\|\boldsymbol{X}\right\|^r\right)^{1/r} \leq \left(\mathbb{E}\left\|\boldsymbol{X}\right\|^p\right)^{1/p} \tag{B.20}$$

Markov's Inequality (standard form). For any random vector $\boldsymbol{x}$ and non-negative function $g(\boldsymbol{x}) \geq 0$,

$$\Pr(g(\boldsymbol{x}) > \alpha) \leq \alpha^{-1}\mathbb{E}g(\boldsymbol{x}). \tag{B.21}$$

Markov's Inequality (strong form). For any random vector $\boldsymbol{x}$ and non-negative function $g(\boldsymbol{x}) \geq 0$,

$$\Pr(g(\boldsymbol{x}) > \alpha) \leq \alpha^{-1}\mathbb{E}\left(g\left(\boldsymbol{x}\right)1\left(g(\boldsymbol{x}) > \alpha\right)\right). \tag{B.22}$$

Chebyshev's Inequality. For any random variable x,

$$\Pr(|x - \mathbb{E}x| > \alpha) \leq \frac{\operatorname{var}(x)}{\alpha^2}. \tag{B.23}$$

Proof of Jensen's Inequality (B.12). 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 (B.12), but using conditional expectations. The conditional expectations exist since $\mathbb{E}\left\|\boldsymbol{y}\right\| < \infty$ and $\mathbb{E}\left\|g\left(\boldsymbol{y}\right)\right\| < \infty$. ■

Proof of Conditional Expectation Inequality. As the function $|u|^r$ is convex for $r \geq 1$, the Conditional Jensen's inequality implies

$$\left|\mathbb{E}(y \mid \boldsymbol{x})\right|^r \leq \mathbb{E}\left(|y|^r \mid \boldsymbol{x}\right).$$

Taking unconditional expectations and the law of iterated expectations, we obtain

$$\mathbb{E}\left|\mathbb{E}(y \mid \boldsymbol{x})\right|^r \leq \mathbb{E}\mathbb{E}\left(|y|^r \mid \boldsymbol{x}\right) = \mathbb{E}\left|y\right|^r < \infty$$

as required. ■

Proof of Expectation Inequality. By the Triangle inequality, for $\lambda \in [0,1]$,

$$\left\|\lambda\boldsymbol{U}_1 + (1-\lambda)\boldsymbol{U}_2\right\| \leq \lambda\left\|\boldsymbol{U}_1\right\| + (1-\lambda)\left\|\boldsymbol{U}_2\right\|$$

which shows that the matrix norm $g(\boldsymbol{U}) = \|\boldsymbol{U}\|$ is convex. Applying Jensen's Inequality (B.12) we find (B.15). ■

Proof of Hölder's Inequality. Since $\frac{1}{p}+\frac{1}{q} = 1$ an application of Jensen's Inequality (A.8) 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.20), $\|\boldsymbol{X}'\boldsymbol{Y}\| \leq \|\boldsymbol{X}\|\,\|\boldsymbol{Y}\|$. 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.16). ■

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

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

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

$$(\mathbb{E}\|\boldsymbol{X}\|^r)^{p/r} \leq \mathbb{E}(\|\boldsymbol{X}\|^r)^{p/r} = \mathbb{E}\|\boldsymbol{X}\|^p.$$

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.21), and then Hölder's Inequality to the two expectations

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{X}+\boldsymbol{Y}\|^p &= \mathbb{E}\left(\|\boldsymbol{X}+\boldsymbol{Y}\|\,\|\boldsymbol{X}+\boldsymbol{Y}\|^{p-1}\right) \\
&\leq \mathbb{E}\left(\|\boldsymbol{X}\|\,\|\boldsymbol{X}+\boldsymbol{Y}\|^{p-1}\right) + \mathbb{E}\left(\|\boldsymbol{Y}\|\,\|\boldsymbol{X}+\boldsymbol{Y}\|^{p-1}\right) \\
&\leq (\mathbb{E}\|\boldsymbol{X}\|^p)^{1/p}\,\mathbb{E}\left(\|\boldsymbol{X}+\boldsymbol{Y}\|^{q(p-1)}\right)^{1/q} \\
&\quad + (\mathbb{E}\|\boldsymbol{Y}\|^p)^{1/p}\,\mathbb{E}\left(\|\boldsymbol{X}+\boldsymbol{Y}\|^{q(p-1)}\right)^{1/q} \\
&= \left((\mathbb{E}\|\boldsymbol{X}\|^p)^{1/p} + (\mathbb{E}\|\boldsymbol{Y}\|^p)^{1/p}\right)\mathbb{E}(\|\boldsymbol{X}+\boldsymbol{Y}\|^p)^{(p-1)/p}
\end{aligned}$$

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 $\mathbb{E}\left(\|\mathbf{X} + \mathbf{Y}\|^p\right)^{(p-1)/p}$, we obtain (B.19). ■

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.22). 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.21). ■

Proof of Chebyshev's Inequality. Define $y = (x - \mathbb{E}x)^2$ and note that $\mathbb{E}y = \operatorname{var}(x)$. The events $\{|x - \mathbb{E}x| > \alpha\}$ and $\left\{y > \alpha^2\right\}$ are equal, so by an application Markov's inequality we find

$$\Pr(|x - \mathbb{E}x| > \alpha) = \Pr(y > \alpha^2) \leq \alpha^{-2}\mathbb{E}\left(y\right) = \alpha^{-2} \operatorname{var}(x)$$

as stated. ■

B.11 Maximum Likelihood

In this section we provide a brief review of the asymptotic theory of maximum likelihood estimation.

When the density of $\boldsymbol{y}_i$ is $f(\boldsymbol{y} \mid \boldsymbol{\theta})$ where F is a known distribution function and $\boldsymbol{\theta} \in \Theta$ is an unknown $m \times 1$ vector, we say that the distribution is **parametric** and that $\boldsymbol{\theta}$ is the **parameter** of the distribution F. The space Θ is the set of permissible value for $\boldsymbol{\theta}$. In this setting the **method of maximum likelihood** is an appropriate technique for estimation and inference on $\boldsymbol{\theta}$. We let $\boldsymbol{\theta}$ denote a generic value of the parameter and let $\boldsymbol{\theta}_0$ denote its true value.

The joint density of a random sample $(\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$ is

$$f_n\left(\boldsymbol{y}_1, ..., \boldsymbol{y}_n \mid \boldsymbol{\theta}\right) = \prod_{i=1}^n f\left(\boldsymbol{y}_i \mid \boldsymbol{\theta}\right).$$

The **likelihood** of the sample is this joint density evaluated at the observed sample values, viewed as a function of $\boldsymbol{\theta}$. The **log-likelihood** function is its natural logarithm

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^n \log f\left(\boldsymbol{y}_i \mid \boldsymbol{\theta}\right).$$

The **likelihood score** is the derivative of the log-likelihood, evaluated at the true parameter value.

$$\boldsymbol{S}_i = \frac{\partial}{\partial\boldsymbol{\theta}} \log f\left(\boldsymbol{y}_i \mid \boldsymbol{\theta}_0\right).$$

We also define the **Hessian**

$$\boldsymbol{\mathcal{H}} = -\mathbb{E}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \log f\left(\boldsymbol{y}_i \mid \boldsymbol{\theta}_0\right) \tag{B.24}$$

and the **outer product** matrix

$$\boldsymbol{\Omega} = \mathbb{E}\left(\boldsymbol{S}_i \boldsymbol{S}_i'\right). \tag{B.25}$$

We now present three important features of the likelihood.

Theorem B.11.1

$$\left.\frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E} \log f(\boldsymbol{y} \mid \boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0} \tag{B.26}$$

$$\mathbb{E}\boldsymbol{S}_i = \mathbf{0} \tag{B.27}$$

and

$$\boldsymbol{\mathcal{H}} = \boldsymbol{\Omega} \equiv \boldsymbol{\mathcal{I}} \tag{B.28}$$

The matrix $\boldsymbol{\mathcal{I}}$ is called the **information**, and the equality (B.28) is called the **information matrix equality**.

The **maximum likelihood estimator** (**MLE**) $\hat{\boldsymbol{\theta}}$ is the parameter value which maximizes the likelihood (equivalently, which maximizes the log-likelihood). We can write this as

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmax}} \log L(\boldsymbol{\theta}). \tag{B.29}$$

In some simple cases, we can find an explicit expression for $\hat{\boldsymbol{\theta}}$ as a function of the data, but these cases are rare. More typically, the MLE $\hat{\boldsymbol{\theta}}$ must be found by numerical methods.

To understand why the MLE $\hat{\boldsymbol{\theta}}$ is a natural estimator for the parameter $\boldsymbol{\theta}$ observe that the standardized log-likelihood is a sample average and an estimator of $\mathbb{E} \log f(\boldsymbol{y}_i \mid \boldsymbol{\theta})$:

$$\frac{1}{n} \log L(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log f(\boldsymbol{y}_i \mid \boldsymbol{\theta}) \xrightarrow{p} \mathbb{E} \log f(\boldsymbol{y}_i \mid \boldsymbol{\theta}).$$

As the MLE $\hat{\boldsymbol{\theta}}$ maximizes the left-hand-side, we can see that it is an estimator of the maximizer of the right-hand-side. The first-order condition for the latter problem is

$$0 = \frac{\partial}{\partial \boldsymbol{\theta}} \mathbb{E} \log f(\boldsymbol{y}_i \mid \boldsymbol{\theta})$$

which holds at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ by (B.26). This suggests that $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}_0$. In. fact, under conventional regularity conditions, $\hat{\boldsymbol{\theta}}$ is consistent, $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ as $n \to \infty$. Furthermore, we can derive its asymptotic distribution.

Theorem B.11.2 *Under regularity conditions,* $\sqrt{n}\left(\hat{\theta} - \theta_0\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{\mathcal{I}}^{-1}\right)$.

We omit the regularity conditions for Theorem B.11.2, but the result holds quite broadly for models which are smooth functions of the parameters. Theorem B.11.2 gives the general form for the asymptotic distribution of the MLE. A famous result shows that the asymptotic variance is the smallest possible.

Theorem B.11.3 ***Cramer-Rao Lower Bound.*** *If $\widetilde{\boldsymbol{\theta}}$ is an unbiased regular estimator of θ, then* $\operatorname{var}(\widetilde{\boldsymbol{\theta}}) \geq (n\boldsymbol{\mathcal{I}})^{-}$.

The Cramer-Rao Theorem shows that the finite sample variance of an unbiased estimator is bounded below by $(n\boldsymbol{\mathcal{I}})^{-1}$. This means that the asymptotic variance of the standardized estimator $\sqrt{n}\left(\widetilde{\boldsymbol{\theta}} - \theta_0\right)$ is bounded below by $\boldsymbol{\mathcal{I}}^{-1}$. In other words, the best possible asymptotic variance among all (regular) estimators is $\boldsymbol{\mathcal{I}}^{-1}$. An estimator is called asymptotically efficient if its asymptotic variance equals this lower bound. Theorem B.11.2 shows that the MLE has this asymptotic variance, and is thus asymptotically efficient.

Theorem B.11.4 *The MLE is asymptotically efficient in the sense that its asymptotic variance equals the Cramer-Rao Lower Bound.*

Theorem B.11.4 gives a strong endorsement for the MLE in parametric models.

Finally, consider functions of parameters. If $\boldsymbol{\psi} = \boldsymbol{g}(\boldsymbol{\theta})$ then the MLE of $\boldsymbol{\psi}$ is $\widehat{\boldsymbol{\psi}} = \boldsymbol{g}(\widehat{\boldsymbol{\theta}})$. This is because maximization (e.g. (B.29)) is unaffected by parameterization and transformation. Applying the Delta Method to Theorem B.11.2 we conclude that

$$\sqrt{n}\left(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}\right) \simeq \boldsymbol{G}'\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}, \boldsymbol{G}'\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{G}\right) \tag{B.30}$$

where $\boldsymbol{G} = \frac{\partial}{\partial \boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{\theta}_0)'$. By Theorem B.11.4, $\widehat{\boldsymbol{\psi}}$ is an asymptotically efficient estimator for $\boldsymbol{\psi}$ since it is the MLE. The asymptotic variance $\boldsymbol{G}'\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{G}$ is the Cramer-Rao lower bound for estimation of $\boldsymbol{\psi}$.

Theorem B.11.5 *The Cramer-Rao lower bound for $\boldsymbol{\psi} = \boldsymbol{g}(\boldsymbol{\theta})$ is $\boldsymbol{G}'\boldsymbol{\mathcal{I}}^{-1}\boldsymbol{G}$, and the MLE $\widehat{\boldsymbol{\psi}} = \boldsymbol{g}(\widehat{\boldsymbol{\theta}})$ is asymptotically efficient.*

Proof of Theorem B.11.1. To see (B.26),

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}}\mathbb{E}\log f\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right)\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= \frac{\partial}{\partial \boldsymbol{\theta}}\int \log f\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right) f\left(\boldsymbol{y} \mid \boldsymbol{\theta}_0\right) d\boldsymbol{y}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} f\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right) \frac{f\left(\boldsymbol{y} \mid \boldsymbol{\theta}_0\right)}{f\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right)} d\boldsymbol{y}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}}\int f\left(\boldsymbol{y} \mid \boldsymbol{\theta}\right) d\boldsymbol{y}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} 1\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \mathbf{0}.
\end{aligned}$$

Equation (B.27) follows by exchanging integration and differentiation

$$\mathbb{E}\frac{\partial}{\partial \boldsymbol{\theta}}\log f\left(\boldsymbol{y} \mid \boldsymbol{\theta}_0\right) = \frac{\partial}{\partial \boldsymbol{\theta}}\mathbb{E}\log f\left(\boldsymbol{y} \mid \boldsymbol{\theta}_0\right) = \mathbf{0}.$$

Similarly, we can show that

$$\mathbb{E}\left(\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)}{f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)}\right)=\mathbf{0}.$$

By direction computation,

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)&=\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)}{f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)}-\frac{\frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)\frac{\partial}{\partial\boldsymbol{\theta}}f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)'}{f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)^2}\\&=\frac{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)}{f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)}-\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\boldsymbol{y}\mid\boldsymbol{\theta}_0\right)'.\end{aligned}$$

Taking expectations yields (B.28). ■

Proof of Theorem B.11.2 Taking the first-order condition for maximization of $\log L(\boldsymbol{\theta})$, and making a first-order Taylor series expansion,

$$\begin{aligned}0&=\left.\frac{\partial}{\partial\boldsymbol{\theta}}\log L(\boldsymbol{\theta})\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\\&=\sum_{i=1}^n\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\boldsymbol{y}_i\mid\hat{\boldsymbol{\theta}}\right)\\&=\sum_{i=1}^n\frac{\partial}{\partial\boldsymbol{\theta}}\log f\left(\boldsymbol{y}_i\mid\boldsymbol{\theta}_0\right)+\sum_{i=1}^n\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(\boldsymbol{y}_i\mid\boldsymbol{\theta}_n\right)\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right),\end{aligned}$$

where $\boldsymbol{\theta}_n$ lies on a line segment joining $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. (Technically, the specific value of $\boldsymbol{\theta}_n$ varies by row in this expansion.) Rewriting this equation, we find

$$\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=\left(-\sum_{i=1}^n\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(\boldsymbol{y}_i\mid\boldsymbol{\theta}_n\right)\right)^{-1}\left(\sum_{i=1}^n\boldsymbol{S}_i\right)$$

where $\boldsymbol{S}_i$ are the likelihood scores. Since the score $\boldsymbol{S}_i$ is mean-zero (B.27) with covariance matrix $\boldsymbol{\Omega}$ (equation B.25) an application of the CLT yields

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n\boldsymbol{S}_i\xrightarrow{d}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right).$$

The analysis of the sample Hessian is somewhat more complicated due to the presence of $\boldsymbol{\theta}_n$. Let $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})=-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(\boldsymbol{y}_i,\boldsymbol{\theta}\right)$. If it is continuous in $\boldsymbol{\theta}$, then since $\boldsymbol{\theta}_n\xrightarrow{p}\boldsymbol{\theta}_0$ it follows that $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_n)\xrightarrow{p}\boldsymbol{\mathcal{H}}$ and so

$$\begin{aligned}-\frac{1}{n}\sum_{i=1}^n\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_n\right)&=\frac{1}{n}\sum_{i=1}^n\left(-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\log f\left(\boldsymbol{y}_i,\boldsymbol{\theta}_n\right)-\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_n)\right)+\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_n)\\&\xrightarrow{p}\boldsymbol{\mathcal{H}}\end{aligned}$$

by an application of a uniform WLLN. (By uniform, we mean that the WLLN holds uniformly over the parameter value. This requires the second derivative to be a smooth function of the parameter.)

Together,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\xrightarrow{d}\boldsymbol{\mathcal{H}}^{-1}\mathrm{N}\left(\mathbf{0},\boldsymbol{\Omega}\right)=\mathrm{N}\left(\mathbf{0},\boldsymbol{\mathcal{H}}^{-1}\boldsymbol{\Omega}\boldsymbol{\mathcal{H}}^{-1}\right)=\mathrm{N}\left(\mathbf{0},\boldsymbol{\mathcal{I}}^{-1}\right),$$

the final equality using Theorem B.11.1 . ■

Proof of Theorem B.11.3. Let $\boldsymbol{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_n)$ be the sample, and set

$$\boldsymbol{S} = \frac{\partial}{\partial \boldsymbol{\theta}} \log f_n (\boldsymbol{Y}, \boldsymbol{\theta}_0) = \sum_{i=1}^{n} \boldsymbol{S}_i$$

which by Theorem (B.11.1) has mean zero and variance $n\boldsymbol{\mathcal{I}}$. 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 $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} = \mathbb{E}\widetilde{\boldsymbol{\theta}} = \int \widetilde{\boldsymbol{\theta}}(\boldsymbol{Y}) f(\boldsymbol{Y}, \boldsymbol{\theta})\, d\boldsymbol{Y}.$$

Differentiating with respect to θ and evaluating at θ_0 yields

$$\begin{aligned} \boldsymbol{I} &= \int \widetilde{\boldsymbol{\theta}}(\boldsymbol{Y}) \frac{\partial}{\partial \boldsymbol{\theta}'} f(\boldsymbol{Y}, \boldsymbol{\theta})\, d\boldsymbol{Y} \\ &= \int \widetilde{\boldsymbol{\theta}}(\boldsymbol{Y}) \frac{\partial}{\partial \boldsymbol{\theta}'} \log f(\boldsymbol{Y}, \boldsymbol{\theta}) f(\boldsymbol{Y}, \boldsymbol{\theta}_0)\, d\boldsymbol{Y} \\ &= \mathbb{E}\left(\widetilde{\boldsymbol{\theta}}\boldsymbol{S}'\right) \\ &= \mathbb{E}\left(\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\boldsymbol{S}'\right) \end{aligned}$$

the final equality since $\mathbb{E}(\boldsymbol{S}) = 0$

By the matrix Cauchy-Schwarz inequality (B.18), $\mathbb{E}\left(\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\boldsymbol{S}'\right) = \boldsymbol{I}$, and $\operatorname{var}(\boldsymbol{S}) = \mathbb{E}(\boldsymbol{S}\boldsymbol{S}') = n\boldsymbol{\mathcal{I}}$,

$$\begin{aligned} \operatorname{var}\left(\widetilde{\boldsymbol{\theta}}\right) &= \mathbb{E}\left(\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)'\right) \\ &\geq \mathbb{E}\left(\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)\boldsymbol{S}'\right) \mathbb{E}(\boldsymbol{S}\boldsymbol{S}')^{-} \mathbb{E}\left(\boldsymbol{S}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)'\right) \\ &= \mathbb{E}(\boldsymbol{S}\boldsymbol{S}')^{-} \\ &= (n\boldsymbol{\mathcal{I}})^{-} \end{aligned}$$

as stated. ■

Appendix C

Numerical Optimization

Many econometric estimators are defined by an optimization problem of the form

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, Q(\boldsymbol{\theta}) \tag{C.1}$$

where the parameter is $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^m$ and the criterion function is $Q(\boldsymbol{\theta}) : \boldsymbol{\Theta} \to \mathbb{R}$. For example NLLS, GLS, MLE and GMM estimators take this form. In most cases, $Q(\boldsymbol{\theta})$ can be computed for given $\boldsymbol{\theta}$, but $\hat{\boldsymbol{\theta}}$ is not available in closed form. In this case, numerical methods are required to obtain $\hat{\boldsymbol{\theta}}$.

C.1 Grid Search

Many optimization problems are either one dimensional ($m = 1$) or involve one-dimensional optimization as a sub-problem (for example, a line search). In this context grid search may be employed.

Grid Search. Let $\Theta = [a, b]$ be an interval. Pick some $\varepsilon > 0$ and set $G = (b - a)/\varepsilon$ to be the number of gridpoints. Construct an equally spaced grid on the region $[a, b]$ with G gridpoints, which is $\{\boldsymbol{\theta}(j) = a + j(b - a)/G : j = 0, ..., G\}$. At each point evaluate the criterion function and find the gridpoint which yields the smallest value of the criterion, which is $\boldsymbol{\theta}(\hat{\jmath})$ where $\hat{\jmath} = \operatorname{argmin}_{0\le j\le G} Q(\boldsymbol{\theta}(j))$. This value $\boldsymbol{\theta}\,(\hat{\jmath})$ is the gridpoint estimate of $\hat{\boldsymbol{\theta}}$**.** If the grid is sufficiently fine to capture small oscillations in $Q(\boldsymbol{\theta})$, the approximation error is bounded by ε, that is, $\left|\boldsymbol{\theta}(\hat{\jmath}) - \hat{\boldsymbol{\theta}}\right| \le \varepsilon$. Plots of $Q(\boldsymbol{\theta}(j))$ against $\boldsymbol{\theta}(j)$ can help diagnose errors in grid selection. This method is quite robust but potentially costly.

Two-Step Grid Search. The gridsearch method can be refined by a two-step execution. For an error bound of ε pick G so that $G^2 = (b - a)/\varepsilon$ For the first step define an equally spaced grid on the region $[a, b]$ with G gridpoints, which is $\{\boldsymbol{\theta}(j) = a + j(b - a)/G : j = 0, ..., G\}$. At each point evaluate the criterion function and let $\hat{\jmath} = \operatorname{argmin}_{0\le j\le G} Q(\boldsymbol{\theta}(j))$. For the second step define an equally spaced grid on $[\boldsymbol{\theta}(\hat{\jmath} - 1), \boldsymbol{\theta}(\hat{\jmath} + 1)]$ with G gridpoints, which is $\{\boldsymbol{\theta}'(k) = \boldsymbol{\theta}(\hat{\jmath} - 1) + 2k(b - a)/G^2 : k = 0, ..., G\}$. Let $\hat{k} = \operatorname{argmin}_{0\le k\le G} Q(\boldsymbol{\theta}'(k))$. The estimate of $\hat{\boldsymbol{\theta}}$ is $\boldsymbol{\theta}\left(\hat{k}\right)$. The advantage of the two-step method over a one-step grid search is that the number of function evaluations has been reduced from $(b-a)/\varepsilon$ to $2\sqrt{(b - a)/\varepsilon}$ which can be substantial. The disadvantage is that if the function $Q(\boldsymbol{\theta})$ is irregular, the first-step grid may not bracket $\hat{\boldsymbol{\theta}}$ which thus would be missed.

C.2 Gradient Methods

Gradient Methods are iterative methods which produce a sequence $\boldsymbol{\theta}_i : i = 1, 2, ...$ which are designed to converge to $\hat{\boldsymbol{\theta}}$. All require the choice of a starting value $\boldsymbol{\theta}_1$, and all require the computation of the **gradient** of $Q(\boldsymbol{\theta})$

$$\boldsymbol{g}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} Q(\boldsymbol{\theta})$$

and some require the **Hessian**

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} Q(\boldsymbol{\theta}).$$

If the functions $\boldsymbol{g}(\boldsymbol{\theta})$ and $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta})$ are not analytically available, they can be calculated numerically. Take the $j'th$ element of $\boldsymbol{g}(\boldsymbol{\theta})$. Let δ_j be the $j'th$ unit vector (zeros everywhere except for a one in the $j'th$ row). Then for ε small

$$g_j(\boldsymbol{\theta}) \simeq \frac{Q(\boldsymbol{\theta} + \delta_j \varepsilon) - Q(\boldsymbol{\theta})}{\varepsilon}.$$

Similarly,

$$g_{jk}(\boldsymbol{\theta}) \simeq \frac{Q(\boldsymbol{\theta} + \delta_j \varepsilon + \delta_k \varepsilon) - Q(\boldsymbol{\theta} + \delta_k \varepsilon) - Q(\boldsymbol{\theta} + \delta_j \varepsilon) + Q(\boldsymbol{\theta})}{\varepsilon^2}$$

In many cases, numerical derivatives can work well but can be computationally costly relative to analytic derivatives. In some cases, however, numerical derivatives can be quite unstable.

Most gradient methods are a variant of **Newton's method** which is based on a quadratic approximation. By a Taylor's expansion for $\boldsymbol{\theta}$ close to $\hat{\boldsymbol{\theta}}$

$$0 = \boldsymbol{g}(\hat{\boldsymbol{\theta}}) \simeq \boldsymbol{g}(\boldsymbol{\theta}) + \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}) \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right)$$

which implies

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta})^{-1} \boldsymbol{g}(\boldsymbol{\theta}).$$

This suggests the iteration rule

$$\hat{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i)^{-1} \boldsymbol{g}(\boldsymbol{\theta}_i).$$

where

One problem with Newton's method is that it will send the iterations in the wrong direction if $\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i)$ is not positive definite. One modification to prevent this possibility is quadratic hill-climbing which sets

$$\hat{\boldsymbol{\theta}}_{i+1} = \boldsymbol{\theta}_i - \left(\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i) + \alpha_i \boldsymbol{I}_m \right)^{-1} \boldsymbol{g}(\boldsymbol{\theta}_i).$$

where α_i is set just above the smallest eigenvalue of $\boldsymbol{H}(\boldsymbol{\theta}_i)$ if $\boldsymbol{H}(\boldsymbol{\theta})$ is not positive definite.

Another productive modification is to add a scalar **steplength** λ_i. In this case the iteration rule takes the form

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{D}_i \boldsymbol{g}_i \lambda_i \tag{C.2}$$

where $\boldsymbol{g}_i = \boldsymbol{g}(\boldsymbol{\theta}_i)$ and $\boldsymbol{D}_i = \boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i)^{-1}$ for Newton's method and $D_i = \left(\boldsymbol{\mathcal{H}}(\boldsymbol{\theta}_i) + \alpha_i \boldsymbol{I}_m \right)^{-1}$ for quadratic hill-climbing.

Allowing the steplength to be a free parameter allows for a line search, a one-dimensional optimization. To pick λ_i write the criterion function as a function of λ

$$Q(\lambda) = Q(\boldsymbol{\theta}_i + \boldsymbol{D}_i \boldsymbol{g}_i \lambda)$$

a one-dimensional optimization problem. There are two common methods to perform a line search. A **quadratic approximation** evaluates the first and second derivatives of $Q(\lambda)$ with respect to λ, and picks λ_i as the value minimizing this approximation. The **half-step** method considers the sequence $\lambda = 1,\ 1/2,\ 1/4,\ 1/8,\ \ldots$. Each value in the sequence is considered and the criterion $Q(\boldsymbol{\theta}_i + \boldsymbol{D}_i \boldsymbol{g}_i \lambda)$ evaluated. If the criterion has improved over $Q(\boldsymbol{\theta}_i)$, use this value, otherwise move to the next element in the sequence.

Newton's method does not perform well if $Q(\boldsymbol{\theta})$ is irregular, and it can be quite computationally costly if $\boldsymbol{H}(\boldsymbol{\theta})$ is not analytically available. These problems have motivated alternative choices for the weight matrix D_i. These methods are called **Quasi-Newton** methods. Two popular methods are do to Davidson-Fletcher-Powell (DFP) and Broyden-Fletcher-Goldfarb-Shanno (BFGS).

Let

$$\begin{aligned}\Delta \boldsymbol{g}_i &= \boldsymbol{g}_i - \boldsymbol{g}_{i-1} \\ \Delta \boldsymbol{\theta}_i &= \boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\end{aligned}$$

and . The DFP method sets

$$\boldsymbol{D}_i = \boldsymbol{D}_{i-1} + \frac{\Delta\boldsymbol{\theta}_i \Delta\boldsymbol{\theta}_i'}{\Delta\boldsymbol{\theta}_i' \Delta\boldsymbol{g}_i} + \frac{\boldsymbol{D}_{i-1}\Delta\boldsymbol{g}_i \Delta\boldsymbol{g}_i' \boldsymbol{D}_{i-1}}{\Delta\boldsymbol{g}_i' \boldsymbol{D}_{i-1}\Delta\boldsymbol{g}_i}.$$

The BFGS methods sets

$$\boldsymbol{D}_i = \boldsymbol{D}_{i-1} + \frac{\Delta\boldsymbol{\theta}_i \Delta\boldsymbol{\theta}_i'}{\Delta\boldsymbol{\theta}_i' \Delta\boldsymbol{g}_i} - \frac{\Delta\boldsymbol{\theta}_i \Delta\boldsymbol{\theta}_i'}{\left(\Delta\boldsymbol{\theta}_i' \Delta\boldsymbol{g}_i\right)^2} \Delta g_i' \boldsymbol{D}_{i-1} \Delta\boldsymbol{g}_i + \frac{\Delta\boldsymbol{\theta}_i \Delta\boldsymbol{g}_i' \boldsymbol{D}_{i-1}}{\Delta\boldsymbol{\theta}_i' \Delta\boldsymbol{g}_i} + \frac{\boldsymbol{D}_{i-1}\Delta\boldsymbol{g}_i \Delta\boldsymbol{\theta}_i'}{\Delta\boldsymbol{\theta}_i' \Delta\boldsymbol{g}_i}.$$

For any of the gradient methods, the iterations continue until the sequence has converged in some sense. This can be defined by examining whether $\left|\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i-1}\right|$, $\left|Q\left(\boldsymbol{\theta}_i\right) - Q\left(\boldsymbol{\theta}_{i-1}\right)\right|$ or $\left|g(\boldsymbol{\theta}_i)\right|$ has become small.

C.3 Derivative-Free Methods

All gradient methods can be quite poor in locating the global minimum when $Q(\boldsymbol{\theta})$ has several local minima. Furthermore, the methods are not well defined when $Q(\boldsymbol{\theta})$ is non-differentiable. In these cases, alternative optimization methods are required. One example is the **simplex method** of Nelder-Mead (1965).

A more recent innovation is the method of **simulated annealing (SA).** For a review see Goffe, Ferrier, and Rodgers (1994). The SA method is a sophisticated random search. Like the gradient methods, it relies on an iterative sequence. At each iteration, a random variable is drawn and added to the current value of the parameter. If the resulting criterion is decreased, this new value is accepted. If the criterion is increased, it may still be accepted depending on the extent of the increase and another randomization. The latter property is needed to keep the algorithm from selecting a local minimum. As the iterations continue, the variance of the random innovations is shrunk. The SA algorithm stops when a large number of iterations is unable to improve the criterion. The SA method has been found to be successful at locating global minima. The downside is that it can take considerable computer time to execute.

Bibliography

[1] Abadir, Karim M. and Jan R. Magnus (2005): *Matrix Algebra*, Cambridge University Press.

[2] Aitken, A.C. (1935): "On least squares and linear combinations of observations," *Proceedings of the Royal Statistical Society*, 55, 42-48.

[3] 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.*

[4] 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.

[5] Andrews, Donald W. K. (1988): "Laws of large numbers for dependent non-identically distributed random variables,' *Econometric Theory,* 4, 458-467.

[6] Andrews, Donald W. K. (1991), "Asymptotic normality of series estimators for nonparameric and semiparametric regression models," *Econometrica*, 59, 307-345.

[7] Andrews, Donald W. K. (1993), "Tests for parameter instability and structural change with unknown change point," *Econometrica*, 61, 821-8516.

[8] Andrews, Donald W. K. and Moshe Buchinsky: (2000): "A three-step method for choosing the number of bootstrap replications," *Econometrica*, 68, 23-51.

[9] Andrews, Donald W. K. and Werner Ploberger (1994): "Optimal tests when a nuisance parameter is present only under the alternative," *Econometrica,* 62, 1383-1414.

[10] Ash, Robert B. (1972): *Real Analysis and Probability*, Academic Press.

[11] Basmann, R. L. (1957): "A generalized classical method of linear estimation of coefficients in a structural equation," *Econometrica,* 25, 77-83.

[12] Bekker, P.A. (1994): "Alternative approximations to the distributions of instrumental variable estimators, *Econometrica,* 62, 657-681.

[13] Billingsley, Patrick (1968): *Convergence of Probability Measures.* New York: Wiley.

[14] Billingsley, Patrick (1995): *Probability and Measure,* 3rd Edition, New York: Wiley.

[15] Bose, A. (1988): "Edgeworth correction by bootstrap in autoregressions," *Annals of Statistics*, 16, 1709-1722.

[16] Box, George E. P. and Dennis R. Cox, (1964). "An analysis of transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211-252.

[17] 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.

[18] Brown, B. W. and Whitney K. Newey (2002): "GMM, efficient bootstrapping, and improved inference ," *Journal of Business and Economic Statistics.*

[19] 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.

[20] 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.

[21] Casella, George and Roger L. Berger (2002): *Statistical Inference*, 2nd Edition, Duxbury Press.

[22] Chamberlain, Gary (1987): "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Econometrics*, 34, 305-334.

[23] Choi, In and Peter C.B. Phillips (1992): "Asymptotic and finite sample distribution theory for IV estimators and tests in partially identified structural equations," *Journal of Econometrics*, 51, 113-150.

[24] Chow, G.C. (1960): "Tests of equality between sets of coefficients in two linear regressions," *Econometrica*, 28, 591-603.

[25] Cragg, John (1992): "Quasi-Aitken Estimation for Heterskedasticity of Unknown Form" *Journal of Econometrics*, 54, 179-201.

[26] Davidson, James (1994): *Stochastic Limit Theory: An Introduction for Econometricians.* Oxford: Oxford University Press.

[27] Davison, A.C. and D.V. Hinkley (1997): *Bootstrap Methods and their Application.* Cambridge University Press.

[28] 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.

[29] Donald Stephen G. and Whitney K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69, 1161-1191.

[30] Dufour, J.M. (1997): "Some impossibility theorems in econometrics with applications to structural and dynamic models," *Econometrica*, 65, 1365-1387.

[31] Efron, Bradley (1979): "Bootstrap methods: Another look at the jackknife," *Annals of Statistics*, 7, 1-26.

[32] Efron, Bradley (1982): *The Jackknife, the Bootstrap, and Other Resampling Plans.* Society for Industrial and Applied Mathematics.

[33] Efron, Bradley and R.J. Tibshirani (1993): *An Introduction to the Bootstrap*, New York: Chapman-Hall.

[34] Eicker, F. (1963): "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34, 447-456.

[35] Engle, Robert F. and Clive W. J. Granger (1987): "Co-integration and error correction: Representation, estimation and testing," *Econometrica*, 55, 251-276.

[36] Frisch, Ragnar (1933): "Editorial," *Econometrica*, 1, 1-4.

[37] Frisch, Ragnar and F. Waugh (1933): "Partial time regressions as compared with individual trends," *Econometrica*, 1, 387-401.

[38] Gallant, A. Ronald and D.W. Nychka (1987): "Seminonparametric maximum likelihood estimation," *Econometrica*, 55, 363-390.

[39] Gallant, A. Ronald and Halbert White (1988): *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. New York: Basil Blackwell.

[40] Galton, Francis (1886): "Regression Towards Mediocrity in Hereditary Stature," The Journal of the Anthropological Institute of Great Britain and Ireland, 15, 246-263.

[41] Goldberger, Arthur S. (1964): *Econometric Theory*, Wiley.

[42] Goldberger, Arthur S. (1968): *Topics in Regression Analysis*, Macmillan

[43] Goldberger, Arthur S. (1991): *A Course in Econometrics*. Cambridge: Harvard University Press.

[44] Goffe, W.L., G.D. Ferrier and J. Rogers (1994): "Global optimization of statistical functions with simulated annealing," *Journal of Econometrics*, 60, 65-99.

[45] Gosset, William S. (a.k.a. "Student") (1908): "The probable error of a mean," *Biometrika*, 6, 1-25.

[46] Gauss, K.F. (1809): "Theoria motus corporum coelestium," in *Werke*, Vol. VII, 240-254.

[47] Granger, Clive W. J. (1969): "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37, 424-438.

[48] Granger, Clive W. J. (1981): "Some properties of time series data and their use in econometric specification," *Journal of Econometrics*, 16, 121-130.

[49] Granger, Clive W. J. and Timo Teräsvirta (1993): *Modelling Nonlinear Economic Relationships*, Oxford University Press, Oxford.

[50] Gregory, A. and M. Veall (1985): "On formulating Wald tests of nonlinear restrictions," *Econometrica*, 53, 1465-1468,

[51] Haavelmo, T. (1944): "The probability approach in econometrics," *Econometrica*, supplement, 12.

[52] Hall, A. R. (2000): "Covariance matrix estimation and the power of the overidentifying restrictions test," *Econometrica*, 68, 1517-1527,

[53] Hall, P. (1992): *The Bootstrap and Edgeworth Expansion*, New York: Springer-Verlag.

[54] Hall, P. (1994): "Methodology and theory for the bootstrap," *Handbook of Econometrics, Vol. IV*, eds. R.F. Engle and D.L. McFadden. New York: Elsevier Science.

[55] Hall, P. and J.L. Horowitz (1996): "Bootstrap critical values for tests based on Generalized-Method-of-Moments estimation," *Econometrica*, 64, 891-916.

[56] Hahn, J. (1996): "A note on bootstrapping generalized method of moments estimators," *Econometric Theory*, 12, 187-197.

[57] Hamilton, James D. (1994) *Time Series Analysis*.

[58] Hansen, Bruce E. (1992): "Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends," *Journal of Econometrics*, 53, 87-121.

[59] Hansen, Bruce E. (1996): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64, 413-430.

[60] 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.

[61] Hansen, Lars Peter (1982): "Large sample properties of generalized method of moments estimators, *Econometrica*, 50, 1029-1054.

[62] 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.

[63] Hausman, J.A. (1978): "Specification tests in econometrics," *Econometrica*, 46, 1251-1271.

[64] Heckman, J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47, 153-161.

[65] 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.

[66] Horowitz, Joel (2001): "The Bootstrap," *Handbook of Econometrics, Vol. 5*, J.J. Heckman and E.E. Leamer, eds., Elsevier Science, 3159-3228.

[67] Imbens, G.W. (1997): "One step estimators for over-identified generalized method of moments models," Review of Economic Studies, 64, 359-383.

[68] Imbens, G.W., R.H. Spady and P. Johnson (1998): "Information theoretic approaches to inference in moment condition models," *Econometrica*, 66, 333-357.

[69] Jarque, C.M. and A.K. Bera (1980): "Efficient tests for normality, homoskedasticity and serial independence of regression residuals, *Economic Letters*, 6, 255-259.

[70] Johansen, S. (1988): "Statistical analysis of cointegrating vectors," *Journal of Economic Dynamics and Control*, 12, 231-254.

[71] Johansen, S. (1991): "Estimation and hypothesis testing of cointegration vectors in the presence of linear trend," *Econometrica*, 59, 1551-1580.

[72] Johansen, S. (1995): *Likelihood-Based Inference in Cointegrated Vector Auto-Regressive Models*, Oxford University Press.

[73] 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.

[74] Kitamura, Y. (2001): "Asymptotic optimality and empirical likelihood for testing moment restrictions," *Econometrica*, 69, 1661-1672.

[75] Kitamura, Y. and M. Stutzer (1997): "An information-theoretic alternative to generalized method of moments," *Econometrica*, 65, 861-874..

[76] Koenker, Roger (2005): *Quantile Regression*. Cambridge University Press.

[77] Kunsch, H.R. (1989): "The jackknife and the bootstrap for general stationary observations," *Annals of Statistics*, 17, 1217-1241.

[78] 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.

[79] Lafontaine, F. and K.J. White (1986): "Obtaining any Wald statistic you want," *Economics Letters*, 21, 35-40.

[80] Lehmann, E.L. and George Casella (1998): *Theory of Point Estimation*, 2nd Edition, Springer.

[81] Lehmann, E.L. and Joseph P. Romano (2005): *Testing Statistical Hypotheses*, 3rd Edition, Springer.

[82] Lindeberg, Jarl Waldemar, (1922): "Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung," *Mathematische Zeitschrift*, 15, 211-225.

[83] Li, Qi and Jeffrey Racine (2007) *Nonparametric Econometrics*.

[84] Lovell, M.C. (1963): "Seasonal adjustment of economic time series," *Journal of the American Statistical Association*, 58, 993-1010.

[85] 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.

[86] MacKinnon, James G. and Halbert White (1985): "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305-325.

[87] Magnus, J. R., and H. Neudecker (1988): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, New York: John Wiley and Sons.

[88] Mann, H.B. and A. Wald (1943). "On stochastic limit and order relationships," *The Annals of Mathematical Statistics* 14, 217–226.

[89] Muirhead, R.J. (1982): *Aspects of Multivariate Statistical Theory.* New York: Wiley.

[90] Nelder, J. and R. Mead (1965): "A simplex method for function minimization," *Computer Journal*, 7, 308-313.

[91] 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.

[92] Newey, Whitney K. (1990): "Semiparametric efficiency bounds," *Journal of Applied Econometrics*, 5, 99-135.

[93] Newey, Whitney K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147-168.

[94] 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.

[95] Newey, Whitney K. and Kenneth D. West (1987): "Hypothesis testing with efficient method of moments estimation," *International Economic Review*, 28, 777-787.

[96] Owen, Art B. (1988): "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 237-249.

[97] Owen, Art B. (2001): *Empirical Likelihood.* New York: Chapman & Hall.

[98] Park, Joon Y. and Peter C. B. Phillips (1988): "On the formulation of Wald tests of nonlinear restrictions," *Econometrica*, 56, 1065-1083,

[99] Phillips, Peter C.B. (1989): "Partially identified econometric models," *Econometric Theory*, 5, 181-240.

[100] Phillips, Peter C.B. and Sam Ouliaris (1990): "Asymptotic properties of residual based tests for cointegration," *Econometrica*, 58, 165-193.

[101] Politis, D.N. and J.P. Romano (1996): "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303-1313.

[102] Potscher, B.M. (1991): "Effects of model selection on inference," *Econometric Theory*, 7, 163-185.

[103] Qin, J. and J. Lawless (1994): "Empirical likelihood and general estimating equations," *The Annals of Statistics*, 22, 300-325.

[104] 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.

[105] Rudin, W. (1987): *Real and Complex Analysis*, 3rd edition. New York: McGraw-Hill.

[106] Runge, Carl (1901): "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten," *Zeitschrift für Mathematik und Physik*, 46, 224-243.

[107] Said, S.E. and D.A. Dickey (1984): "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, 71, 599-608.

[108] Secrist, Horace (1933): *The Triumph of Mediocrity in Business.* Evanston: Northwestern University.

[109] Shao, J. and D. Tu (1995): *The Jackknife and Bootstrap.* NY: Springer.

[110] Sargan, J.D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica, 2*6, 393-415.

[111] Shao, Jun (2003): *Mathematical Statistics*, 2nd edition, Springer.

[112] 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.

[113] Shin, Y. (1994): "A residual-based test of the null of cointegration against the alternative of no cointegration," *Econometric Theory*, 10, 91-115.

[114] Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis.* London: Chapman and Hall.

[115] Sims, C.A. (1972): "Money, income and causality," *American Economic Review*, 62, 540-552.

[116] Sims, C.A. (1980): "Macroeconomics and reality," *Econometrica, 48*, 1-48.

[117] Staiger, D. and James H. Stock (1997): "Instrumental variables regression with weak instruments," *Econometrica,* 65, 557-586.

[118] Stock, James H. (1987): "Asymptotic properties of least squares estimators of cointegrating vectors," *Econometrica,* 55, 1035-1056.

[119] Stock, James H. (1991): "Confidence intervals for the largest autoregressive root in U.S. macroeconomic time series," *Journal of Monetary Economics,* 28, 435-460.

[120] Stock, James H. and Jonathan H. Wright (2000): "GMM with weak identification," *Econometrica,* 68, 1055-1096.

[121] Stock, James H. and Mark W. Watson (2010): *Introduction to Econometrics*, 3rd edition, Addison-Wesley.

[122] Stone, Marshall H. (1937): "Applications of the Theory of Boolean Rings to General Topology," *Transactions of the American Mathematical Society*, 41, 375-481.

[123] Stone, Marshall H. (1948): "The Generalized Weierstrass Approximation Theorem," *Mathematics Magazine*, 21, 167-184.

[124] Theil, Henri. (1953): "Repeated least squares applied to complete equation systems," The Hague, Central Planning Bureau, mimeo.

[125] Theil, Henri (1961): *Economic Forecasts and Policy.* Amsterdam: North Holland.

[126] Theil, Henri. (1971): *Principles of Econometrics*, New York: Wiley.

[127] Tobin, James (1958): "Estimation of relationships for limited dependent variables," *Econometrica, 2*6, 24-36.

[128] Tripathi, Gautam (1999): "A matrix extension of the Cauchy-Schwarz inequality," Economics Letters, 63, 1-3.

[129] van der Vaart, A.W. (1998): *Asymptotic Statistics*, Cambridge University Press.

[130] Wald, A. (1943): "Tests of statistical hypotheses concerning several parameters when the number of observations is large," *Transactions of the American Mathematical Society*, 54, 426-482.

[131] Wang, J. and E. Zivot (1998): "Inference on structural parameters in instrumental variables regression with weak instruments," *Econometrica,* 66, 1389-1404.

[132] Weierstrass, K. (1885): "Über die analytische Darstellbarkeit sogenannter willkürlicher Functionen einer reellen Veränderlichen," *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*, 1885.

[133] White, Halbert (1980): "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817-838.

[134] White, Halbert (1984): *Asymptotic Theory for Econometricians,* Academic Press.

[135] Wooldridge, Jeffrey M. (2010) *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, MIT Press.

[136] Wooldridge, Jeffrey M. (2009) *Introductory Econometrics: A Modern Approach,* 4th edition, Southwestern

[137] 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.

[138] Zhang, Fuzhen and Qingling Zhang (2006): "Eigenvalue inequalities for matrix product," *IEEE Transactions on Automatic Control*, 51, 1506-1509.