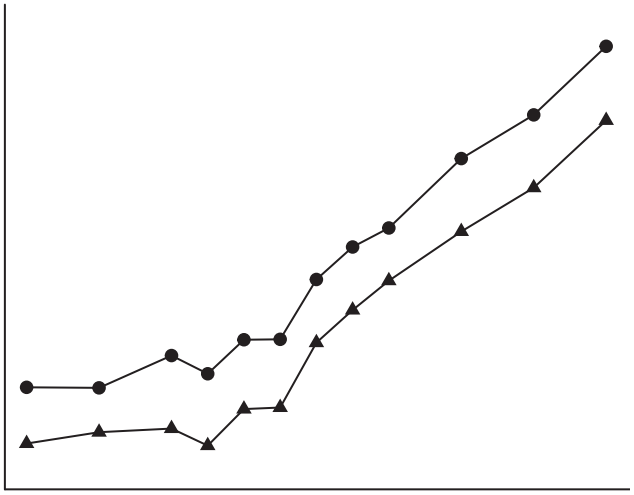
ECONOMETRICS



BRUCE E. HANSEN

Contents

Preface

This textbook is the second in a two-part series covering the core material typically taught in a one-year Ph.D. course in econometrics. . IThe sequence is:

1. *Probability and Statistics for Economists* (first volume)
2. *Econometrics* (this volume)

Econometrics assumes that students have a background in 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 Stock and Watson (2019) and Wooldridge (2020). The relevant background in probability theory and mathematical statistics is provided in *Probability and Statistics for Economists*.

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

This textbook contains more material than can be covered in a one-semester course. This is intended to provide instructors flexibility concerning which topics to cover, which to cover in depth, and which to cover briefly. Some material is suitable for second-year PhD instruction. At the University of Wisconsin, where this material was developed, in the first half of the fall semester we cover *Probability and Statistics for Economists*. In the second half of the fall semester we cover Chapters 1-9 of *Econometrics*. In the first half of the spring semester we cover Chapters 10-17, with some chapters covered briefly. In the second half of the spring semester we cover Chapters 18-29, with many details only covered briefly. We revisit much of the latter material in our second-year curriculum, with greater focus on the econometric theory.

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

For further study in econometrics beyond this text, I recommend Davidson (1994) for asymptotic theory, Hamilton (1994) and Kilian and Lütkepohl (2017) for time series methods, Cameron and Trivedi (2005) and Wooldridge (2010) for panel data and discrete response models, and Li and Racine (2007) for nonparametrics and semiparametric econometrics. Beyond these texts, the *Handbook of Econometrics* series provides advanced summaries of contemporary econometric methods and theory.

Alternative PhD-level econometrics textbooks include Theil (1971), Amemiya (1985), Judge, Griffiths, Hill, Lütkepohl, and Lee (1985), Goldberger (1991), Davidson and MacKinnon (1993), Johnston and DiNardo (1997), Davidson (2000), Hayashi (2000), Ruud (2000), Davidson and MacKinnon (2004), Greene (2018), and Magnus (2017). For a focus on applied issues see Angrist and Pischke (2009) and Cunningham (2021).

The end-of-chapter exercises are important parts of the text and are meant to help teach students of econometrics.

Acknowledgements

This book and its companion *Probability and Statistics for Economists* would not have been possible if it were not for the amazing flow of unsolicited advice, corrections, comments, and questions I have received from students, faculty, and other readers over the twenty years I have worked on this project. I have received emails corrections and comments from so many individuals I have completely lost track of the list. So rather than publish an incomplete list, I simply give an honest and thorough *Thank You* to every single one.

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

My most heartfelt thanks goes to my family: Korinna, Zoe, and Nicholas. Without their love and support over these years this project would not have been possible.

100% of the author's royalties will be re-gifted to charitable purposes or graduate student needs.

Notation

Real numbers (elements of the real line $\mathbb{R}$, also called **scalars**) are written using lower case italics such as x.

Vectors (elements of $\mathbb{R}^k$) are typically written by lower case italics such as x, and sometimes using lower case bold italics such as $\boldsymbol{x}$ (for matrix algebra expressions), For example, we write

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

Vectors by default are written as column vectors. The **transpose** of x is the row vector

$$x' = \begin{pmatrix} x_1 & x_2 & \cdots & x_m \end{pmatrix}.$$

There is diversity between fields concerning the choice of notation for the transpose. The notation x' is the most common in econometrics. In statistics and mathematics the notation $x^\top$ is typically used, or occassionally x^t.

Matrices are written using upper case bold italics. For example

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}.$$

Random variables and vectors are written using upper case italics such as X.

We typically use Greek letters such as β, θ, and σ^2 to denote parameters of a probability model. Estimators are typically denoted by putting a hat "^", tilde "~" or bar "-" over the corresponding letter, e.g. $\widehat{\beta}$ and $\widetilde{\beta}$ are estimators of β.

Common Symbols

a	scalar
a or $\boldsymbol{a}$	vector
$\boldsymbol{A}$	matrix
X	random variable or vector
$\mathbb{R}$	real line
$\mathbb{R}_+$	positive real line
$\mathbb{R}^k$	Euclidean k space
$\mathbb{P}[A]$	probability
$\mathbb{P}[A \mid B]$	conditional probability
$F(x)$	cumulative distribution function
$f(x)$	probability density function
$\mathbb{E}[X]$	mathematical expectation
$\mathbb{E}[Y \mid X = x]$, $\mathbb{E}[Y \mid X]$	conditional expectation
$\text{var}[X]$	variance or covariance matrix
$\text{var}[Y \mid X = x]$, $\text{var}[Y \mid X]$	conditional variance
$\text{cov}(X, Y)$	covariance
$\mathscr{P}[Y \mid X = x]$, $\mathscr{P}[Y \mid X]$	best linear predictor
$\text{corr}(X, Y)$	correlation
$\overline{X}$	sample mean
$\widehat{\sigma}^2$	sample variance
s^2	biased-corrected sample variance
$\widehat{\theta}$	estimator
$s\left(\widehat{\theta}\right)$	standard error of estimator
$\lim_{n\to\infty}$	limit
$\text{plim}_{n\to\infty}$	probability limit
$\longrightarrow$	convergence
$\xrightarrow[p]{}$	convergence in probability
$\xrightarrow[d]{}$	convergence in distribution
$L_n(\theta)$	likelihood function
$\ell_n(\theta)$	log-likelihood function
$\mathscr{I}_\theta$	information matrix
$\text{N}(0,1)$	standard normal distribution
$\text{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
χ^2_k	chi-square distribution with k degrees of freedom

$\boldsymbol{I}_n$	$n \times n$ identity matrix
$\mathbf{1}_n$	$n \times 1$ vector of ones
$\operatorname{tr} \boldsymbol{A}$	trace
$\boldsymbol{a}'$ or $\boldsymbol{A}'$	vector or matrix transpose
$\boldsymbol{A}^{-1}$	matrix inverse
$\boldsymbol{A} > 0$	positive definite
$\boldsymbol{A} \geq 0$	positive semi-definite
$\|\boldsymbol{a}\|$	Euclidean norm
$\|\boldsymbol{A}\|$	matrix norm
$\stackrel{\text{def}}{=}$	definitional equality
$\mathbb{1}\{a\}$	indicator function (1 if a is true, else 0)
$\simeq$	approximate equality
$\sim$	is distributed as
$\log(x)$	natural logarithm
$\exp(x)$	exponential function
$\sum_{i=1}^{n}$	summation from $i = 1$ to n

Greek Alphabet

It is common in economics and econometrics to use Greek characters to augment the Latin alphabet. The following table lists the various Greek characters and their pronunciations in English. The second character, when listed, is upper case (except for ϵ which is an alternative script for ε.)

Greek Character	Name	Latin Keyboard Equivalent
α	alpha	a
β	beta	b
γ, Γ	gamma	g
δ, Δ	delta	d
ε, ϵ	epsilon	e
ζ	zeta	z
η	eta	h
θ, Θ	theta	y
ι	iota	i
κ	kappa	k
λ, Λ	lambda	l
μ	mu	m
ν	nu	n
ξ, Ξ	xi	x
π, Π	pi	p
ρ	rho	r
σ, Σ	sigma	s
τ	tau	t
υ	upsilon	u
ϕ, Φ	phi	f
χ	chi	x
ψ, Ψ	psi	c
ω, Ω	omega	w

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 viewpoints, that of statistics, economic theory, and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the *unification* of all three that is powerful. And it is this unification that constitutes econometrics.
>
> Ragnar Frisch, *Econometrica*, (1933), 1, pp. 1-2.

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

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

1.2 The Probability Approach to Econometrics

The unifying methodology of modern econometrics was articulated by Trygve Haavelmo (1911-1999) of Norway, winner of the 1989 Nobel Memorial Prize in Economic Sciences, in his seminal paper "The probability approach in econometrics" (1944). Haavelmo argued that quantitative economic models must necessarily be *probability models* (by which today we would mean *stochastic*). Deterministic models are blatently inconsistent with observed economic quantities, and it is incoherent to apply deterministic models to non-deterministic data. Economic models should be explicitly designed to incorporate randomness; stochastic errors should not be simply added to deterministic models to make them random. Once we acknowledge that an economic model is a probability model, it follows naturally that an appropriate tool way to quantify, estimate, and conduct inferences about the economy is through the powerful theory of mathematical statistics. The appropriate method for a quantitative economic analysis follows from the probabilistic construction of the economic model.

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

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

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

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

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

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

Trygve Haavelmo

The founding ideas of the field of econometrics are largely due to the Norweigen econometrician Trygve Haavelmo (1911-1999). His advocacy of probability models revolutionized the field, and his use of formal mathematical reasoning laid the foundation for subsequent generations. He was awarded the Nobel Memorial Prize in Economic Sciences in 1989.

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

1.3 Econometric Terms

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 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.

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 practice, random variables and vectors are denoted by upper case roman characters such as Y and X. We make an exception for equation errors which we typically denote by the lower case letters e, u, or v.

Real numbers (elements of the real line $\mathbb{R}$, also called **scalars**) are written using lower case italics such as x. Vectors (elements of $\mathbb{R}^k$) are typically also written using lower case italics such as x, or using lower case bold italics such as $\boldsymbol{x}$. We use bold in matrix algebraic expressions for compatibility with matrix notation.

Matrices are written using upper case bold italics such as $\boldsymbol{X}$. Our notation will not make a distinction between random and non-random matrices. Typically we use $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$ to denote random matrices and use $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{W}$ to denote non-random 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. In some contexts we use indices other than i, such as in time series applications where the index t is common. In panel studies we typically use the double index it to refer to individual i at a time period t.

We typically use Greek letters such as β, θ, and σ^2 to denote unknown parameters (scalar or vectors). Parameter matrices are written using upper case Latin boldface, e.g. $\boldsymbol{A}$. Estimators are typically denoted by putting a hat "^", tilde "~", or bar "-" over the corresponding letter, e.g. $\widehat{\beta}$ and $\widetilde{\beta}$ are estimators of β, and $\widehat{\boldsymbol{A}}$ is an estimator of $\boldsymbol{A}$.

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

1.4 Observational Data

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

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

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

1.5 Standard Data Structures

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

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

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

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

Clustered samples are increasing popular in applied economics and are related to panel data. In clustered sampling the observations are grouped into "clusters" which are treated as mutually independent yet allowed to be dependent within the cluster. The major difference with panel data is that clustered

sampling typically does not explicitly model error component structures, nor the dependence within clusters, but rather is concerned with inference which is robust to arbitrary forms of within-cluster correlation.

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

Data Structures

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

As we mentioned above, most of this text will be devoted to cross-sectional data under the assumption of mutually independent observations. By mutual independence we mean that the i^{th} observation (Y_i, X_i) is independent of the j^{th} observation $\left(Y_j, X_j\right)$ for $i \neq j$. In this case we say that the data are **independently distributed**. (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 X_i.)

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

Definition 1.1 The variables (Y_i, X_i) are a **sample** from the distribution F if they are identically distributed with distribution F.

Definition 1.2 The variables (Y_i, X_i) are a **random sample** if they are mutually independent and identically distributed (i.i.d.) across $i = 1, ..., n$.

In the random sampling framework, we think of an individual observation (Y_i, X_i) as a realization from a joint probability distribution $F\left(y, x\right)$ which we 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 because the distribution F is unknown, and the goal of statistical inference is to learn about features of F from the sample. The *assumption* of random sampling provides the mathematical foundation for treating economic statistics with the tools of mathematical statistics.

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

1.6 Econometric Software

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

Stata 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. At many points in this textbook specific Stata estimation methods and commands are described. These commands are valid for Stata version 16.

MATLAB, GAUSS, and OxMetrics 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.

An intermediate choice is R. R has the capabilities of the above high-level matrix programming languages, but also has many built-in statistical environments which can replicate much of the functionality of Stata. R is the dominant programming language in the statistics field, so methods developed in that arena are most commonly available in R. Uniquely, R is open-source, user-contributed, and best of all, completely free! A growing group of econometricians are enthusiastic fans of R.

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.

There are many other packages which are used by econometricians, include Eviews, Gretl, PcGive, Python, Julia, RATS, and SAS.

As the packages described above have distinct advantages many empirical economists use multiple packages. As a student of econometrics you will learn at least one of these packages and probably more than one. My advice is that all students of econometrics should develop a basic level of familiarity with Stata, MATLAB, and R.

1.7 Replication

Scientific research needs to be documented and replicable. For social science research using observational data this requires careful documentation and archiving of the research methods, data manipulations, and coding.

The best practice is as follows. Accompanying each published paper an author should create a complete replication package (set of data files, documentation, and program code files). This package should contain the source (raw) data used for analysis, and code which executes the empirical analysis and other numerical work reported in the paper. In most cases this is a set of programs which may need to be executed sequentially. (For example, there may be an initial program which "cleans" and manipulates the data, and then a second set of programs which estimate the reported models.) The ideal is full documentation and clarity. This package should be posted on the author(s) website, and posted at the journal website when that is an option.

A complicating factor is that many current economic data sets have restricted access and cannot be shared without permission. In these cases the data cannot be posted nor shared. The computed code, however, can and should be posted.

Most journals in economics require authors of published papers to make their datasets generally available. For example:

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:

> It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented and access to the data and code is non-exclusive to the authors. Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to acceptance, information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs.

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.

1.8 Data Files for Textbook

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

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

The data sets currently included are

- `AB1991`
  - Data file from Arellano and Bond (1991)
- `AJR2001`
  - Data file from Acemoglu, Johnson, and Robinson (2001)
- `AK1991`
  - Data file from Angrist and Krueger (1991)
- `AL1999`
  - Data file from Angrist and Lavy (1999)
- `BMN2016`
  - Data file from Bernheim, Meer and Novarro (2016)
- `cps09mar`
  - household survey data extracted from the March 2009 Current Population Survey
- `Card1995`
  - Data file from Card (1995)
- `CHJ2004`
  - Data file from Cox, B. E. Hansen and Jimenez (2004)
- `CK1994`
  - Data file from Card and Krueger (1994)
- `CMR2008`
  - Date file from Card, Mas, and Rothstein (2008)

- `DDK2011`
  - Data file from Duflo, Dupas, and Kremer (2011)
- `DS2004`
  - Data file from DiTella and Schargrodsky (2004)
- `FRED-MD` and `FRED-QD`
  - U.S. monthly and quarterly macroeconomic databases from McCracken and Ng (2015)
- `Invest1993`
  - Data file from Hall and Hall (1993)
- `LM2007`
  - Data file from Ludwig and Miller (2007) and Cattaneo, Titiunik, and Vazquez-Bare (2017)
- `Kilian2009`
  - Data file from Kilian (2009)
- `Koppelman`
  - Data file from Forinash and Koppelman (1993), Koppelman and Wen (2000) and Wen and Koppelman (2001)
- `MRW1992`
  - Data file from Mankiw, Romer, and Weil (1992)
- `Nerlove1963`
  - Data file from Nerlov (1963)
- `PSS2017`
  - Data file from Papageorgiou, Saam, and Schulte (2017)
- `RR2010`
  - Data file from Reinhard and Rogoff (2010)

1.9 Reading the Manuscript

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

To fully understand econometric methods it is necessary to have a mathematical understanding of its mechanics, and this includes the mathematical proofs of the main results. Consequently, this text is self-contained with nearly all results proved with full mathematical rigor. The mathematical development and proofs aim at brevity and conciseness (sometimes described as mathematical elegance), but also at pedagogy. To understand a mathematical proof it is not sufficient to simply *read* the proof, you need to follow it and re-create it for yourself.

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

Key concepts in matrix algebra and a set of useful inequalities are reviewed in Appendices A & B. It may be useful to read or review Appendix A.1-A.11 before starting Chapter 3, and review Appendix B before Chapter 6. It is not necessary to understand all the material in the appendices. They are intended to be reference material and some of the results are not used in this textbook.

Part I

Regression

Chapter 2

Conditional Expectation and Projection

2.1 Introduction

The most commonly applied econometric tool is least squares estimation, also known as **regression**. Least squares is a tool to estimate the 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. This includes a review of probability theory. For a background in intermediate probability theory see Chapters 1-5 of *Probability and Statistics for Economists*.

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**

$$F(y) = \mathbb{P}\left[wage \leq y\right].$$

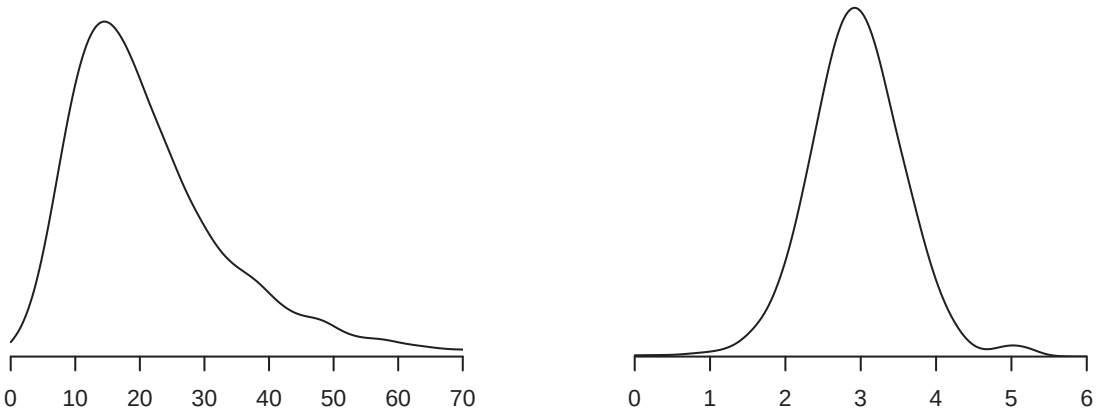
When we say that a person's wage is random we mean that we do not know their wage before it is measured, and we treat observed wage rates as realizations from the distribution F. Treating unobserved wages as random variables and observed wages as realizations is a powerful mathematical abstraction which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

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

$$f(y) = \frac{d}{dy}F(y).$$

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



(a) Wage Density

(b) Log Wage Density

Figure 2.1: Density of Wages and Log Wages

In Figure 2.1(a) we display an estimate[1] of the probability density function 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 distribution F is the unique solution to

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

The median U.S. wage is $19.23. The median is a robust[2] measure of central tendency, but it is tricky to use for many calculations as it is not a linear operator.

The **mean** or **expectation** of a random variable Y with discrete support is

$$\mu = \mathbb{E}[Y] = \sum_{j=1}^{\infty} \tau_j \mathbb{P}\left[Y = \tau_j\right].$$

For a continuous random variable with density $f(y)$ the expectation is

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

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*. An alternative notation which includes both discrete and continuous random variables as special cases is to write the integral as $\int_{-\infty}^{\infty} y dF(y)$.

The expectation is a convenient measure of central tendency because it is a linear operator and arises naturally in many economic models. A disadvantage of the expectation is that it is not robust[3] especially

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

[3] The expectation is sensitive to pertubations in the tails of the distribution.

in the presence of substantial skewness or thick tails, both which are features of the wage distribution as can be seen in Figure 2.1(a). Another way of viewing this is that 64% of workers earn less than the mean wage of $23.90, suggesting that it is incorrect to describe the mean $23.90 as a "typical" wage rate.

In this context it is useful to transform the data by taking the natural logarithm[4]. Figure 2.1(b) shows the density of log hourly wages log(*wage*) for the same population. The density of log wages is 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 better (more robust) measure[5] 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 $F(y)$ is in terms of its **quantiles**. For any $\alpha \in (0,1)$, the α^{th} quantile of the continuous[6] 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 called **percentiles**. E.g. the median is the 50^{th} percentile.

2.3 Conditional Expectation

We saw in Figure 2.1(b) 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. To investigate, we plot in Figure 2.2(a) the densities of log wages for U.S. men and women. We can see that the two wage densities take similar shapes but the density for men is somewhat shifted to the right.

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 expectation** (or **conditional mean**) of log wages given gender. We can write their specific values as

$$\mathbb{E}\left[\log(wage) \mid gender = man\right] = 3.05$$

$$\mathbb{E}\left[\log(wage) \mid gender = woman\right] = 2.81.$$

We call these expectations "conditional" as they are conditioning on a fixed value of the variable *gender*. While you might not think of a person's gender as a random variable, it is random from the viewpoint of econometric analysis. If you randomly select an individual, the gender 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.

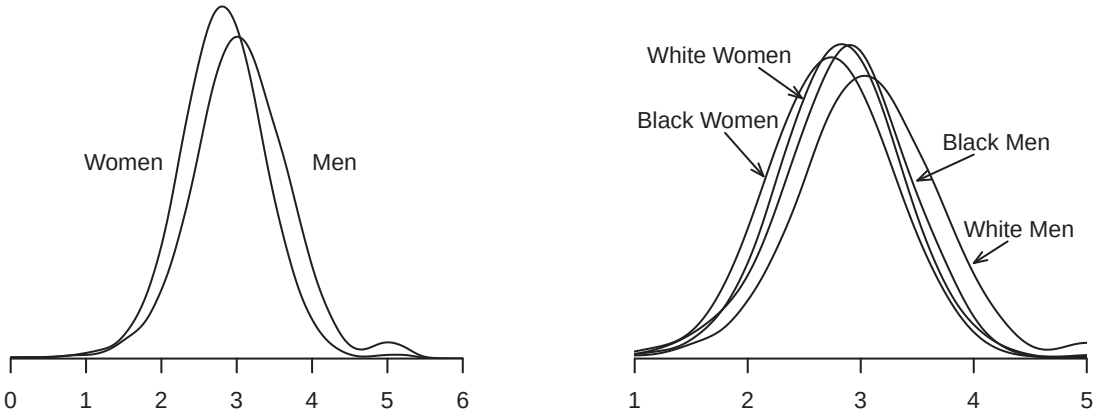
It is important to mention at this point that we in no way attribute causality or interpretation to the difference in the conditional expectation of log wages between men and women. There are multiple potential explanations.

As the two densities in Figure 2.2(a) 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 more carefully. As we mentioned above, the

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

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

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



(a) Women and Men

(b) By Gender and Race

Figure 2.2: Log Wage Density by Gender and Race

primary difference between the two densities appears to be their means. This difference equals

$$\begin{aligned}\mathbb{E}\left[\log(wage) \mid gender = man\right] - \mathbb{E}\left[\log(wage) \mid gender = woman\right] &= 3.05 - 2.81 \\ &= 0.24. \end{aligned} \tag{2.1}$$

A difference in expected log wages of 0.24 is often interpreted as an average 24% difference between the wages of men and women, which is quite substantial. (For a more complete explanation see Section 2.4.)

Consider further splitting the male and female subpopulations by race, dividing the population into whites, Blacks, and other races. We display the log wage density functions of four of these groups in Figure 2.2(b). Again we see that the primary difference between the four density functions is their central tendency.

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

Table 2.1: Mean Log Wages by Gender and Race

	men	women
white	3.07	2.82
Black	2.86	2.73
other	3.03	2.86

Once again we stress that we in no way attribute causality or interpretation to the differences across the entries of the table. The reason why we use these particular sub-populations to illustrate conditional expectation is because differences in economic outcomes between gender and racial groups in the United States (and elsewhere) are widely discussed; part of the role of social science is to carefully document such patterns, and part of its role is to craft models and explanations. Conditional expectations

(by themselves) can help in the documentation and description; conditional expectations by themselves are neither a model nor an explanation.

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

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

and

$$\mathbb{E}\left[\log(wage) \mid gender = woman,\ race = Black\right] = 2.73.$$

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 Logs and Percentages

In this section we want to motivate and clarify the use of the logarithm in regression analysis by making two observations. First, when applied to numbers the difference of logarithms approximately equals the percentage difference. Second, when applied to averages the difference in logarithms approximately equals the percentage difference in the geometric mean. We now explain these ideas and the nature of the approximations involved.

Take two positive numbers a and b. The percentage difference between a and b is

$$p = 100\left(\frac{a-b}{b}\right).$$

Rewriting,

$$\frac{a}{b} = 1 + \frac{p}{100}.$$

Taking natural logarithms,

$$\log a - \log b = \log\left(1 + \frac{p}{100}\right). \tag{2.2}$$

A useful approximation for small x is

$$\log(1+x) \simeq x. \tag{2.3}$$

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

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

The symbol $O(x^2)$ means that the remainder is bounded by Ax^2 as $x \to 0$ for some $A < \infty$. Numerically, the approximation $\log(1+x) \simeq x$ is within 0.001 for $|x| \leq 0.1$, and the approximation error increases with $|x|$.

Applying (2.3) to (2.2) and multiplying by 100 we find

$$p \simeq 100\left(\log a) - \log b\right).$$

This shows that 100 multiplied by the difference in logarithms is approximately the percentage difference. Numerically, the approximation error is less than 0.1 percentage points for $\left|p\right| \leq 10$.

Now consider the difference in the expectation of log transformed random variables. Take two random variables $X_1, X_2 > 0$. Define their geometric means $\theta_1 = \exp\left(\mathbb{E}\left[\log X_1\right]\right)$ and $\theta_2 = \exp\left(\mathbb{E}\left[\log X_2\right]\right)$ and their percentage difference

$$p = 100\left(\frac{\theta_2 - \theta_1}{\theta_1}\right).$$

The difference in the expectation of the log transforms (multiplied by 100) is

$$100\left(\mathbb{E}\left[\log X_2\right] - \mathbb{E}\left[\log X_1\right]\right) = 100\left(\log\theta_2 - \log\theta_1\right) \simeq p$$

the percentage difference between θ_2 and θ_1. In words, the difference between the average of the log transformed variables is (approximately) the percentage difference in the geometric means.

The reason why this latter observation is important is because many econometric equations take the semi-log form

$$\mathbb{E}\left[\log Y \mid group = 1\right] = \mu_1$$
$$\mathbb{E}\left[\log Y \mid group = 2\right] = \mu_2$$

and considerable attention is given to the difference $\mu_1 - \mu_2$. For example, in the previous section we compared the average log wages for men and women and found that the difference is 0.24. In that section we stated that this difference is often interpreted as the average percentage difference. This is not quite right, but is not quite wrong either. What the above calculation shows is that this difference is approximately the percentage difference in the geometric mean. So $\mu_1 - \mu_2$ is an average percentage difference, where "average" refers to geometric rather than arithmetic mean.

To compare different measures of percentage difference see Table 2.2. In the first two columns we report average wages for men and women in the CPS population using four "averages": arithmetic mean, median, geometric mean, and mean log. For both groups the arithmetic mean is higher than the median and geometric mean, and the latter two are similar to one another. This is a common feature of skewed distributions such as the wage distribution. The third column reports the percentage difference between the first two columns (using men's wages as the base). For example, the first entry of 34% states that the mean wage for men is 34% higher than the mean wage for women. The next entries show that the median and geometric mean for men is 26% higher than those for women. The final entry in this column is 100 times the simple difference between the mean log wage, which is 24%. As shown above, the difference in the mean of the log transformation is approximately the percentage difference in the geometric mean, and this approximation is excellent for differences under 10%.

Let's summarize this analysis. It is common to take logarithms of variables and make comparisons between conditional means. We have shown that these differences are measures of the percentage difference in the geometric mean. Thus the common description that the difference between expected log transforms (such as the 0.24 difference between those for men and women's wages) is an approximate percentage difference (e.g. a 24% difference in men's wages relative to women's) is correct, so long as we realize that we are implicitly comparing geometric means.

2.5 Conditional Expectation Function

An important determinant of wages is education. In many empirical studies economists measure educational attainment by the number of years[7] of schooling. We will write this variable as *education*.

[7] 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 *educa-*

Table 2.2: Average Wages and Percentage Differences

	men	women	% Difference
Arithmetic Mean	$26.80	$20.00	34%
Median	$21.14	$16.83	26%
Geometric Mean	$21.03	$16.64	26%
Mean log Wage	3.05	2.81	24%

The conditional expectation of log(*wage*) given *gender*, *race*, and *education* is a single number for each category. For example

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

We display in Figure 2.3 the conditional expectation of log(*wage*) as a function of *education*, separately for (white) men and women. The plot is quite revealing. We see that the conditional expectation is increasing in years of education, but at a different rate for schooling levels above and below nine years. Another striking feature of Figure 2.3 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.

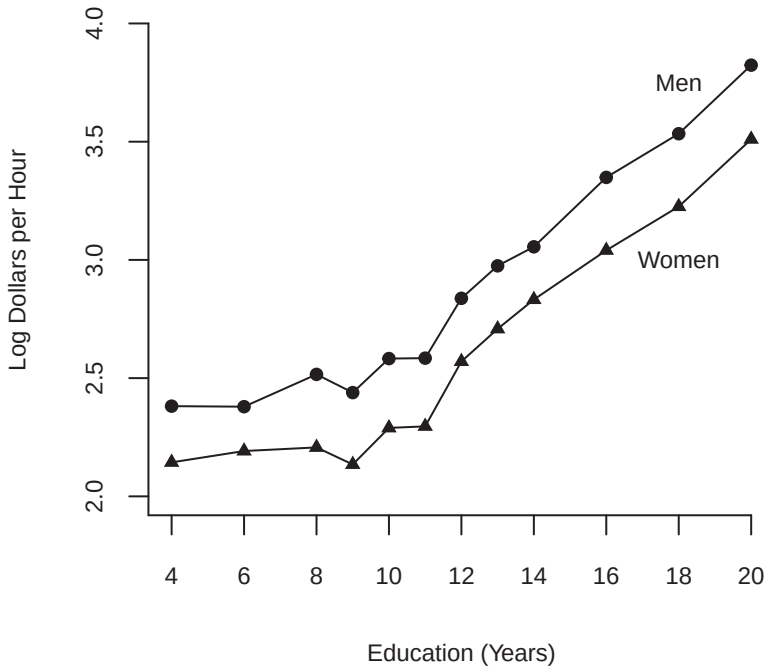


Figure 2.3: Expected Log Wage as a Function of Education

tion=20.

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 *gender*) by the letter X, and multiple conditioning variables (such as *race*, *education* and *gender*) by the subscripted letters $X_1, X_2, ..., X_k$.

Conditional expectations can be written with the generic notation

$$\mathbb{E}[Y \mid X_1 = x_1, X_2 = x_2, ..., X_k = x_k] = m(x_1, x_2, ..., x_k).$$

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 $(X_1, X_2) = (gender, race)$ is given by the six entries of Table 2.1.

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

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}. \tag{2.4}$$

Given this notation, the CEF can be compactly written as

$$\mathbb{E}[Y \mid X = x] = m(x).$$

The CEF $m(x) = \mathbb{E}[Y \mid X = x]$ is a function of $x \in \mathbb{R}^k$. It says: "When X takes the value x then the average value of Y is $m(x)$." Sometimes it is useful to view the CEF as a function of the random variable X. In this case we evaluate the function $m(x)$ at X, and write $m(X)$ or $\mathbb{E}[Y \mid X]$. This is random as it is a function of the random variable X.

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, X) are continuously distributed with a joint density function $f(y, x)$.

As an example, take $Y = \log(wage)$ and $X = experience$, the latter the number of years of potential labor market experience[8]. The contours of their joint density are plotted in Figure 2.4(a) for the population of white men with 12 years of education.

Given the joint density $f(y, x)$ the variable x has the marginal density

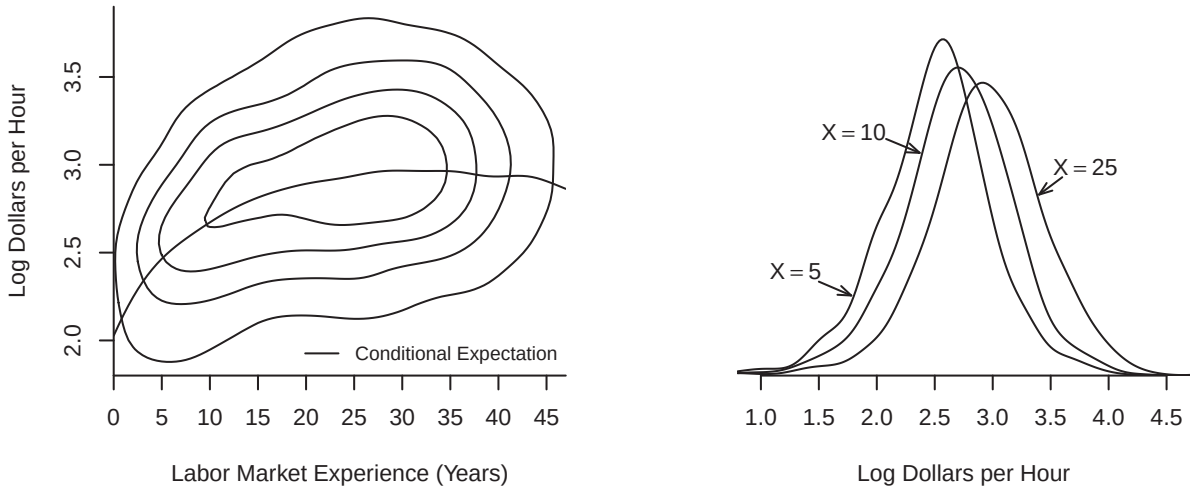
$$f_X(x) = \int_{-\infty}^{\infty} f(y, x)\, dy.$$

For any x such that $f_X(x) > 0$ the conditional density of Y given X is defined as

$$f_{Y|X}(y \mid x) = \frac{f(y, x)}{f_X(x)}. \tag{2.5}$$

The conditional density is a renormalized slice of the joint density $f(y, x)$ holding x fixed. The slice is renormalized (divided by $f_X(x)$ so that it integrates to one) and is thus a density. We can visualize this by slicing the joint density function at a specific value of x parallel with the y-axis. For example, take the density contours in Figure 2.4(a) and slice through the contour plot at a specific value of *experience*, and

[8] As there is no direct measure for experience, we instead define *experience* as *age*−*education*−6



(a) Joint Density of Log Wage and Experience (b) Conditional Density of Log Wage given Experience

Figure 2.4: Log Wage and Experience

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 three levels of *experience* (5, 10, and 25 years), and plot these densities in Figure 2.4(b). We can see that the distribution of wages shifts to the right and becomes more diffuse as experience increases.

The CEF of Y given $X = x$ is the expectation of the conditional density (2.5)

$$m(x) = \mathbb{E}[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)\, dy. \tag{2.6}$$

Intuitively, $m(x)$ is the expectation of Y for the idealized subpopulation where the conditioning variables are fixed at x. When X is continuously distributed this subpopulation is infinitely small.

This definition (2.6) is appropriate when the conditional density (2.5) is well defined. However, Theorem 2.13 in Section 2.31 will show that $m(x)$ can be defined for any random variables (Y, X) so long as $\mathbb{E}|Y| < \infty$.

In Figure 2.4(a) 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 known as the Simple Law.

Theorem 2.1 Simple Law of Iterated Expectations
If $\mathbb{E}|Y| < \infty$ then for any random vector X,

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y].$$

This 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. For discrete X

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \sum_{j=1}^{\infty} \mathbb{E}\left[Y \mid X = x_j\right] \mathbb{P}\left[X = x_j\right].$$

For continuous X

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \int_{\mathbb{R}^k} \mathbb{E}[Y \mid X = x] f_X(x) dx.$$

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

$$\begin{aligned}
&\mathbb{E}\left[\log(wage) \mid gender = man\right] \mathbb{P}\left[gender = man\right] \\
&+ \mathbb{E}\left[\log(wage) \mid gender = woman\right] \mathbb{P}\left[gender = woman\right] \\
&= \mathbb{E}\left[\log(wage)\right].
\end{aligned}$$

Or numerically,

$$3.05 \times 0.57 + 2.81 \times 0.43 = 2.95.$$

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

Theorem 2.2 Law of Iterated Expectations
If $\mathbb{E}|Y| < \infty$ then for any random vectors X_1 and X_2,

$$\mathbb{E}[\mathbb{E}[Y \mid X_1, X_2] \mid X_1] = \mathbb{E}[Y \mid X_1].$$

Notice the way the law is applied. The inner expectation conditions on X_1 and X_2, while the outer expectation conditions only on X_1. The iterated expectation yields the simple answer $\mathbb{E}[Y \mid X_1]$, the expectation conditional on X_1 alone. Sometimes we phrase this as: "The smaller information set wins."

As an example

$$\begin{aligned}
&\mathbb{E}\left[\log(wage) \mid gender = man,\ race = white\right] \mathbb{P}\left[race = white \mid gender = man\right] \\
&+ \mathbb{E}\left[\log(wage) \mid gender = man,\ race = Black\right] \mathbb{P}\left[race = Black \mid gender = man\right] \\
&+ \mathbb{E}\left[\log(wage) \mid gender = man,\ race = other\right] \mathbb{P}\left[race = other \mid gender = man\right] \\
&= \mathbb{E}\left[\log(wage) \mid gender = man\right]
\end{aligned}$$

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 X you can effectively treat it as if it is constant. For example, $\mathbb{E}[X \mid X] = X$ and $\mathbb{E}\left[g(X) \mid X\right] = g(X)$ for any function $g(\cdot)$. The general property is known as the Conditioning Theorem.

Theorem 2.3 Conditioning Theorem
If $\mathbb{E}|Y| < \infty$ then

$$\mathbb{E}\left[g(X)\, Y \mid X\right] = g(X)\,\mathbb{E}[Y \mid X]. \tag{2.7}$$

If in addition $\mathbb{E}\left|g(X)\right| < \infty$ then

$$\mathbb{E}\left[g(X)\, Y\right] = \mathbb{E}\left[g(X)\,\mathbb{E}[Y \mid X]\right]. \tag{2.8}$$

The proofs of Theorems 2.1, 2.2 and 2.3 are given in Section 2.33.

2.8 CEF Error

The CEF error e is defined as the difference between Y and the CEF evaluated at X:

$$e = Y - m(X).$$

By construction, this yields the formula

$$Y = m(X) + e. \tag{2.9}$$

In (2.9) it is useful to understand that the error e is derived from the joint distribution of (Y, X), and so its properties are derived from this construction.

Many authors in econometrics denote the CEF error using the Greek letter ε. I do not follow this convention because the error e is a random variable similar to Y and X, and it is typical to use Latin characters for random variables.

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

$$\begin{aligned}\mathbb{E}[e \mid X] &= \mathbb{E}[(Y - m(X)) \mid X] \\ &= \mathbb{E}[Y \mid X] - \mathbb{E}[m(X) \mid X] \\ &= m(X) - m(X) = 0.\end{aligned}$$

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

$$\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[e \mid X]] = \mathbb{E}[0] = 0.$$

We state this and some other results formally.

Theorem 2.4 Properties of the CEF error
If $\mathbb{E}|Y| < \infty$ then

1. $\mathbb{E}[e \mid X] = 0$.
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 $h(x)$ such that $\mathbb{E}|h(X)\, e| < \infty$ then $\mathbb{E}[h(X)\, e] = 0$.

The proof of the third result is deferred to Section 2.33. The fourth result, whose proof is left to Exercise 2.3, implies that e is uncorrelated with any function of the regressors.

The equations

$$\begin{aligned} Y &= m(X) + e \\ \mathbb{E}[e \mid X] &= 0 \end{aligned}$$

together imply that $m(X)$ is the CEF of Y given X. It is important to understand that this is not a restriction. These equations hold true by definition.

The condition $\mathbb{E}[e \mid X] = 0$ is implied by the definition of e as the difference between Y and the CEF $m(X)$. The equation $\mathbb{E}[e \mid X] = 0$ is sometimes called a conditional mean restriction, because 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 X. However, it does not imply that the distribution of e is independent of X. Sometimes the assumption "e is independent of X" is added as a convenient simplification, but it is not generic feature of the conditional mean. Typically and generally, e and X are jointly dependent even though the conditional mean of e is zero.

As an example, the contours of the joint density of the regression error e and *experience* are plotted in Figure 2.5 for the same population as Figure 2.4. Notice that the shape of the conditional distribution varies with the level of *experience*.

Figure 2.5: Joint Density of Regression Error and Experience

As a simple example of a case where X and e are mean independent yet dependent let $e = Xu$ where X and u 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 X. In this case $m(X) = \mathbb{E}[Y] = \mu$, the unconditional expectation 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 2.4.3 implies the following simple but useful result.

Theorem 2.5 If $\mathbb{E}\left[Y^2\right] < \infty$ then $\sigma^2 < \infty$.

We can call σ^2 the **regression variance** or the variance of the regression error. The magnitude of σ^2 measures the amount of variation in Y which is not "explained" or accounted for in the conditional expectation $\mathbb{E}[Y \mid X]$.

The regression variance depends on the regressors X. Consider two regressions

$$\begin{aligned} Y &= \mathbb{E}[Y \mid X_1] + e_1 \\ Y &= \mathbb{E}[Y \mid X_1, X_2] + e_2. \end{aligned}$$

We write the two errors distinctly as e_1 and e_2 as they are different – changing the conditioning information changes the conditional expectation 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 expectation $\mathbb{E}[Y \mid X]$ as the "explained portion" of Y. The remainder $e = Y - \mathbb{E}[Y \mid X]$ 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 amount of information always decreases the variance of the unexplained portion.

Theorem 2.6 If $\mathbb{E}\left[Y^2\right] < \infty$ then

$$\operatorname{var}[Y] \geq \operatorname{var}[Y - \mathbb{E}[Y \mid X_1]] \geq \operatorname{var}[Y - \mathbb{E}[Y \mid X_1, X_2]].$$

Theorem 2.6 says that the variance of the difference between Y and its conditional expectation (weakly) decreases whenever an additional variable is added to the conditioning information.

The proof of Theorem 2.6 is given in Section 2.33.

2.11 Best Predictor

Suppose that given a random vector X we want to predict or forecast Y. We can write any predictor as a function $g(X)$ of X. The (ex-post) prediction error is the realized difference $Y - g(X)$. A non-stochastic measure of the magnitude of the prediction error is the expectation of its square

$$\mathbb{E}\left[\left(Y - g(X)\right)^2\right]. \tag{2.10}$$

We can define the best predictor as the function $g(X)$ which minimizes (2.10). What function is the best predictor? It turns out that the answer is the CEF $m(X)$. This holds regardless of the joint distribution of (Y, X).

To see this, note that the mean squared error of a predictor $g(X)$ is

$$\begin{aligned}
\mathbb{E}\left[\left(Y - g(X)\right)^2\right] &= \mathbb{E}\left[\left(e + m(X) - g(X)\right)^2\right] \\
&= \mathbb{E}\left[e^2\right] + 2\mathbb{E}\left[e\left(m(X) - g(X)\right)\right] + \mathbb{E}\left[\left(m(X) - g(X)\right)^2\right] \\
&= \mathbb{E}\left[e^2\right] + \mathbb{E}\left[\left(m(X) - g(X)\right)^2\right] \\
&\geq \mathbb{E}\left[e^2\right] \\
&= \mathbb{E}\left[(Y - m(X))^2\right].
\end{aligned}$$

The first equality makes the substitution $Y = m(X) + e$ and the third equality uses Theorem 2.4.4. The right-hand-side after the third equality is minimized by setting $g(X) = m(X)$, yielding the inequality in the fourth line. The minimum is finite under the assumption $\mathbb{E}\left[Y^2\right] < \infty$ as shown by Theorem 2.5.

We state this formally in the following result.

Theorem 2.7 Conditional Expectation as Best Predictor
If $\mathbb{E}\left[Y^2\right] < \infty$, then for any predictor $g(X)$,

$$\mathbb{E}\left[\left(Y - g(X)\right)^2\right] \geq \mathbb{E}\left[(Y - m(X))^2\right]$$

where $m(X) = \mathbb{E}[Y \mid X]$.

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 2.7 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 Y.

Definition 2.1 If $\mathbb{E}\left[Y^2\right] < \infty$, the **conditional variance** of Y given $X = x$ is

$$\sigma^2(x) = \operatorname{var}[Y \mid X = x] = \mathbb{E}\left[(Y - \mathbb{E}[Y \mid X = x])^2 \mid X = x\right].$$

The conditional variance treated as a random variable is $\operatorname{var}[Y \mid X] = \sigma^2(X)$.

The conditional variance is distinct from the **unconditional** variance $\operatorname{var}[Y]$. The difference is that the conditional variance is a function of the conditioning variables. 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 2.2 If $\mathbb{E}\left[e^2\right] < \infty$, the **conditional variance** of the regression error e given $X = x$ is

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

The conditional variance of e treated as a random variable is $\operatorname{var}[e \mid X] = \sigma^2(X)$.

Again, the conditional variance $\sigma^2(x)$ is distinct from the unconditional variance σ^2. The conditional variance is a function of the regressors, the unconditional variance is not. Generally, $\sigma^2(x)$ is a non-trivial function of x and can take any form subject to the restriction that it is non-negative. One way to think about $\sigma^2(x)$ is that it is the conditional mean of e^2 given X. Notice as well that $\sigma^2(x) = \operatorname{var}[Y \mid X = x]$ so it is equivalently the conditional variance of the dependent variable.

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

As an example of how the conditional variance depends on observables, compare the conditional log wage densities for men and women displayed in Figure 2.2. 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 variance is related to the conditional variance by the following identity.

Theorem 2.8 If $\mathbb{E}\left[Y^2\right] < \infty$ then

$$\operatorname{var}[Y] = \mathbb{E}[\operatorname{var}[Y \mid X]] + \operatorname{var}[\mathbb{E}[Y \mid X]].$$

See Theorem 4.14 of *Probability and Statistics for Economists*. Theorem 2.8 decomposes the unconditional variance into what are sometimes called the "within group variance" and the "across group variance". For example, if X is education level, then the first term is the expected variance of the conditional expectation by education level. The second term is the variance after controlling for education.

The regression error has a conditional mean of zero, so its unconditional error variance equals the expected conditional variance, or equivalently can be found by the law of iterated expectations.

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

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

Given the conditional variance we can define a rescaled error

$$u = \frac{e}{\sigma(X)}. \tag{2.11}$$

We calculate that since $\sigma(X)$ is a function of X

$$\mathbb{E}[u \mid X] = \mathbb{E}\left[\frac{e}{\sigma(X)} \,\middle|\, X\right] = \frac{1}{\sigma(X)}\mathbb{E}[e \mid X] = 0$$

and

$$\operatorname{var}[u \mid X] = \mathbb{E}\left[u^2 \mid X\right] = \mathbb{E}\left[\frac{e^2}{\sigma^2(X)} \,\middle|\, X\right] = \frac{1}{\sigma^2(X)}\mathbb{E}\left[e^2 \mid X\right] = \frac{\sigma^2(X)}{\sigma^2(X)} = 1.$$

Thus u has a conditional expectation of zero and a conditional variance of 1.

Notice that (2.11) can be rewritten as

$$e = \sigma(X)u.$$

and substituting this for e in the CEF equation (2.9), we find that

$$Y = m(X) + \sigma(X)u.$$

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

Many econometric studies focus on the conditional expectation $m(x)$ and either ignore the conditional variance $\sigma^2(x)$, treat it as a constant $\sigma^2(x) = \sigma^2$, 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 is 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 $\sigma^2(x)$ is a constant and independent of x. This is called **homoskedasticity**.

Definition 2.3 The error is **homoskedastic** if $\sigma^2(x) = \sigma^2$ does not depend on x.

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

Definition 2.4 The error is **heteroskedastic** if $\sigma^2(x)$ depends on 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 X. So when we talk about the variance as a function of the regressors we are talking about the conditional variance $\sigma^2(x)$.

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 $\sigma^2(x)$ depends on observables.

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

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

Heteroskedastic or Heteroscedastic?

The spelling of the words *homoskedastic* and *heteroskedastic* have been somewhat controversial. Early econometrics textbooks were split, with some using a "c" as in *heteroscedastic* and some "k" as in *heteroskedastic*. McCulloch (1985) pointed out that the word is derived from Greek roots. $o\mu o\iota o\varsigma$ means "same". $\varepsilon\tau\varepsilon\rho o$ means "other" or "different". $\sigma\kappa\varepsilon\delta\alpha\nu\nu\upsilon\mu\iota$ means "to scatter". Since the proper transliteration of the Greek letter κ in $\sigma\kappa\varepsilon\delta\alpha\nu\nu\upsilon\mu\iota$ is "k", this implies that the correct English spelling of the two words is with a "k" as in *homoskedastic* and *heteroskedastic*.

2.14 Regression Derivative

One way to interpret the CEF $m(x) = \mathbb{E}[Y \mid X = x]$ is in terms of how marginal changes in the regressors X imply changes in the conditional expectation 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

$$\nabla_1 m(x) = \begin{cases} \dfrac{\partial}{\partial x_1} m(x_1, ..., x_k), & \text{if } X_1 \text{ is continuous} \\ \\ m(1, x_2, ..., x_k) - m(0, x_2, ..., x_k), & \text{if } X_1 \text{ is binary.} \end{cases}$$

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

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

When all elements of X are continuous, then we have the simplification $\nabla m(x) = \dfrac{\partial}{\partial x} m(x)$, 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 expectation does not literally hold *all else* constant. It only holds constant the variables included in the conditional expectation. This means that the regression derivative depends on which regressors are included. For example, in a regression of wages on education, experience, race and gender, the regression derivative with respect to education shows the marginal effect of education on expected wages, holding constant experience, race, and gender. 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 $\nabla m(x)$ is the change in the actual value of Y only if the error e is unaffected by the change in the regressor 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 $m(x) = \mathbb{E}[Y \mid X = x]$ is linear in x. In this case we can write the mean equation as

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

Notationally it is convenient to write this as a simple function of the vector x. An easy way to do so is to augment the regressor vector 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[9] of the vector x is $x_k = 1$. Thus (2.4) has been redefined as the $k \times 1$ vector

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_{k-1} \\ 1 \end{pmatrix}. \tag{2.12}$$

With this redefinition, the CEF is

$$m(x) = x_1\beta_1 + x_2\beta_2 + \cdots + \beta_k = x'\beta \tag{2.13}$$

where

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

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 X.

In the linear CEF model the regression derivative is simply the coefficient vector. That is $\nabla m(x) = \beta$. 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

$$Y = X'\beta + e$$
$$\mathbb{E}[e \mid X] = 0$$

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

Homoskedastic Linear CEF Model

$$Y = X'\beta + e$$
$$\mathbb{E}[e \mid X] = 0$$
$$\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$$

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

2.16 Linear CEF with Nonlinear Effects

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

For example, suppose we have two scalar variables X_1 and X_2. The CEF could take the quadratic form

$$m(x_1, x_2) = x_1\beta_1 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6. \tag{2.14}$$

This equation is quadratic in the regressors (x_1, x_2) yet linear in the coefficients $\beta = (\beta_1, ..., \beta_6)'$. We still call (2.14) a linear CEF because it is a linear function of the coefficients. At the same time, it has nonlinear effects because it is nonlinear in the underlying variables x_1 and x_2. The key is to understand that (2.14) is quadratic in the variables (x_1, x_2) yet linear in the coefficients β.

To simplify the expression we define the transformations $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1x_2$, and $x_6 = 1$, and redefine the regressor vector as $x = (x_1, ..., x_6)'$. With this redefinition, $m(x_1, x_2) = x'\beta$ which is linear in β. For most econometric purposes (estimation and inference on β) the linearity in β is all that is important.

An exception is in the analysis of regression derivatives. In nonlinear equations such as (2.14) 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 (2.14), the regression derivatives are not a simple coefficient, but are functions of several coefficients plus the levels of (x_1, x_2). Consequently it is difficult to interpret the coefficients individually. It is more useful to interpret them as a group.

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

2.17 Linear CEF with Dummy Variables

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

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

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

$$\mathbb{E}\left[Y \mid gender\right] = \begin{cases} \mu_0 & \text{if} \quad gender = man \\ \\ \mu_1 & \text{if} \quad gender = woman. \end{cases}$$

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

$$X_1 = \begin{cases} 0 & \text{if} \quad gender = man \\ 1 & \text{if} \quad gender = woman. \end{cases} \tag{2.15}$$

Given this notation we write the conditional expectation as a linear function of the dummy variable X_1. Thus $\mathbb{E}[Y \mid X_1] = \beta_1 X_1 + \beta_2$ 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 expectation of Y for the $X_1 = 0$ subpopulation (men) and the slope β_1 is equal to the difference in the conditional expectations between the two subpopulations.

Alternatively, we could have defined X_1 as

$$X_1 = \begin{cases} 1 & \text{if} \quad \textit{gender} = \textit{man} \\ 0 & \text{if} \quad \textit{gender} = \textit{woman}. \end{cases} \tag{2.16}$$

In this case, the regression intercept is the expectation 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 "gender" does not help distinguish between definitions (2.15) and (2.16). Instead, it is better to label X_1 as "women" or "female" if definition (2.15) is used, or as "men" or "male" if (2.16) 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 expectation 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, 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 gender on expected log wages for unmarried wage earners, the coefficient β_2 as the effect of marriage on expected log wages for men wage earners, and the coefficient β_3 as the difference between the effects of marriage on expected log wages among women and among men. Alternatively, it can also be interpreted as the difference between the effects of gender 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 $(X_1, X_2, X_1 X_2)$. To put the model in the framework of Section 2.15 we define the regressor $X_3 = X_1 X_2$ and the regressor vector as

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ 1 \end{pmatrix}.$$

So while we started with two dummy variables, the number of regressors (including the intercept) is four.

If there are three 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

$$\mathbb{E}[Y \mid X_1, X_2, X_3] = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_{7X1} X_2 X_3 + \beta_8$$

which has eight regressors including the intercept.

In general, if there are p dummy variables $X_1, ..., X_p$ then the CEF $\mathbb{E}\left[Y \mid X_1, X_2, ..., X_p\right]$ 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. A linear regression model which includes all 2^p binary interactions is called a **saturated dummy variable regression model**. It is a complete model of the conditional expectation. In contrast, a model with no interactions equals

$$\mathbb{E}\left[Y \mid X_1, X_2, ..., X_p\right] = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \beta_p.$$

This has $p+1$ coefficients instead of 2^p.

We started this section by saying that the conditional expectation 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 expectation 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 expectation 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 expectation $m(X) = \mathbb{E}[Y \mid X]$ is the best predictor of Y among all functions of X, its functional form is typically unknown. In particular, the linear CEF model is empirically unlikely to be accurate unless X is discrete and low-dimensional so all interactions are included. Consequently, in most cases it is more realistic to view the linear specification (2.13) as an approximation. In this section we derive a specific approximation with a simple interpretation.

Theorem 2.7 showed that the conditional expectation $m(X)$ 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 2.1

1. $\mathbb{E}\left[Y^2\right] < \infty$.
2. $\mathbb{E}\|X\|^2 < \infty$.
3. $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is positive definite.

In Assumption 2.1.2 we use $\|x\| = \left(x'x\right)^{1/2}$ to denote the Euclidean length of the vector x.

The first two parts of Assumption 2.1 imply that the variables Y and 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 $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ are linearly independent and that the matrix is invertible.

A linear predictor for Y is a function $X'\beta$ for some $\beta \in \mathbb{R}^k$. The mean squared prediction error is

$$S(\beta) = \mathbb{E}\left[\left(Y - X'\beta\right)^2\right]. \tag{2.17}$$

The **best linear predictor** of Y given X, written $\mathscr{P}[Y \mid X]$, is found by selecting the β which minimizes $S(\beta)$.

Definition 2.5 The **Best Linear Predictor** of Y given X is

$$\mathscr{P}[Y \mid X] = X'\beta$$

where β minimizes the mean squared prediction error

$$S(\beta) = \mathbb{E}\left[\left(Y - X'\beta\right)^2\right].$$

The minimizer

$$\beta = \operatorname*{argmin}_{b \in \mathbb{R}^k} S(b) \tag{2.18}$$

is called the **Linear Projection Coefficient**.

We now calculate an explicit expression for its value. The mean squared prediction error (2.17) can be written out as a quadratic function of β:

$$S(\beta) = \mathbb{E}\left[Y^2\right] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}\left[XX'\right]\beta. \tag{2.19}$$

The quadratic structure of $S(\beta)$ means that we can solve explicitly for the minimizer. The first-order condition for minimization (from Appendix A.20) is

$$0 = \frac{\partial}{\partial\beta}S(\beta) = -2\mathbb{E}[XY] + 2\mathbb{E}\left[XX'\right]\beta. \tag{2.20}$$

Rewriting (2.20) as

$$2\mathbb{E}[XY] = 2\mathbb{E}\left[XX'\right]\beta$$

and dividing by 2, this equation takes the form

$$\boldsymbol{Q}_{XY} = \boldsymbol{Q}_{XX}\beta \tag{2.21}$$

where $\boldsymbol{Q}_{XY} = \mathbb{E}[XY]$ is $k \times 1$ and $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is $k \times k$. The solution is found by inverting the matrix $\boldsymbol{Q}_{XX}$, and is written

$$\beta = \boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY}$$

or

$$\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY]. \tag{2.22}$$

It is worth taking the time to understand the notation involved in the expression (2.22). $\boldsymbol{Q}_{XX}$ is a $k \times k$ matrix and $\boldsymbol{Q}_{XY}$ is a $k \times 1$ column vector. Therefore, alternative expressions such as $\frac{\mathbb{E}[XY]}{\mathbb{E}[XX']}$ or $\mathbb{E}[XY]\left(\mathbb{E}\left[XX'\right]\right)^{-1}$ are incoherent and incorrect. We also can now see the role of Assumption 2.1.3. It is equivalent to assuming that $\boldsymbol{Q}_{XX}$ has an inverse $\boldsymbol{Q}_{XX}^{-1}$ which is necessary for the solution to the normal equations (2.21) to be unique, and equivalently for (2.22) to be uniquely defined. In the absence of Assumption 2.1.3 there could be multiple solutions to the equation (2.21).

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

$$\mathscr{P}[Y \mid X] = X'\left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY].$$

This expression is also referred to as the **linear projection** of Y on X.

The **projection error** is

$$e = Y - X'\beta. \tag{2.23}$$

This equals the error (2.9) from the regression equation when (and only when) the conditional expectation is linear in X, otherwise they are distinct.

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

$$Y = X'\beta + e. \tag{2.24}$$

In general, we call equation (2.24) or $X'\beta$ the best linear predictor of Y given X, or the linear projection of Y on X. Equation (2.24) is also often called the **regression** of Y on 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

$$\mathbb{E}[Xe] = 0. \tag{2.25}$$

To see this, using the definitions (2.23) and (2.22) and the matrix properties $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}$ and $\boldsymbol{I}\boldsymbol{a} = \boldsymbol{a}$,

$$\begin{aligned}\mathbb{E}[Xe] &= \mathbb{E}\left[X\left(Y - X'\beta\right)\right] \\ &= \mathbb{E}[XY] - \mathbb{E}\left[XX'\right]\left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY] \\ &= 0\end{aligned} \tag{2.26}$$

as claimed.

Equation (2.25) is a set of k equations, one for each regressor. In other words, (2.25) is equivalent to

$$\mathbb{E}\left[X_j e\right] = 0 \tag{2.27}$$

for $j = 1, ..., k$. As in (2.12), the regressor vector X typically contains a constant, e.g. $X_k = 1$. In this case (2.27) for $j = k$ is the same as

$$\mathbb{E}[e] = 0. \tag{2.28}$$

Thus the projection error has a mean of zero when the regressor vector contains a constant. (When X does not have a constant (2.28) 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 because $\operatorname{cov}(X_j, e) = \mathbb{E}\left[X_j e\right] - \mathbb{E}\left[X_j\right]\mathbb{E}[e]$, then (2.27)-(2.28) 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 2.9 Properties of Linear Projection Model
Under Assumption 2.1,

1. The moments $\mathbb{E}\left[XX'\right]$ and $\mathbb{E}[XY]$ exist with finite elements.
2. The linear projection coefficient (2.18) exists, is unique, and equals
$$\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY].$$
3. The best linear predictor of Y given X is
$$\mathscr{P}(Y \mid X) = X'\left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY].$$
4. The projection error $e = Y - X'\beta$ exists. It satisfies $\mathbb{E}\left[e^2\right] < \infty$ and $\mathbb{E}[Xe] = 0$.
5. If X contains an constant, then $\mathbb{E}[e] = 0$.
6. If $\mathbb{E}|Y|^r < \infty$ and $\mathbb{E}\|X\|^r < \infty$ for $r \geq 2$ then $\mathbb{E}|e|^r < \infty$.

A complete proof of Theorem 2.9 is given in Section 2.33.

It is useful to reflect on the generality of Theorem 2.9. The only restriction is Assumption 2.1. Thus for any random variables (Y, X) with finite variances we can define a linear equation (2.24) with the properties listed in Theorem 2.9. Stronger assumptions (such as the linear CEF model) are not necessary. In this sense the linear model (2.24) exists quite generally. However, it is important not to misinterpret the generality of this statement. The linear equation (2.24) 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

$$
\begin{aligned}
Y &= X'\beta + e \\
\mathbb{E}[Xe] &= 0 \\
\beta &= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY]
\end{aligned}
$$

Invertibility and Identification

The linear projection coefficient $\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY]$ exists and is unique as long as the $k \times k$ matrix $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is invertible. The matrix $\boldsymbol{Q}_{XX}$ is often called the **design matrix** as in experimental settings the researcher is able to control $\boldsymbol{Q}_{XX}$ by manipulating the distribution of the regressors X.

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

$$\alpha'\boldsymbol{Q}_{XX}\alpha = \mathbb{E}\left[\alpha'XX'\alpha\right] = \mathbb{E}\left[\left(\alpha'X\right)^2\right] \geq 0$$

so $\boldsymbol{Q}_{XX}$ by construction is positive semi-definite, conventionally written as $\boldsymbol{Q}_{XX} \geq 0$. The assumption that it is positive definite means that this is a strict inequality, $\mathbb{E}\left[\left(\alpha'X\right)^2\right] > 0$. This is conventionally written as $\boldsymbol{Q}_{XX} > 0$. This condition means that there is no non-zero vector α such that $\alpha'X = 0$ identically. Positive definite matrices are invertible. Thus when $\boldsymbol{Q}_{XX} > 0$ then $\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}[XY]$ exists and is uniquely defined. In other words, if we can exclude the possibility that a linear function of X is degenerate, then β is uniquely defined.

Theorem 2.5 shows that the linear projection coefficient β is **identified** (uniquely determined) under Assumption 2.1. The key is invertibility of $\boldsymbol{Q}_{XX}$. Otherwise, there is no unique solution to the equation

$$\boldsymbol{Q}_{XX}\beta = \boldsymbol{Q}_{XY}. \tag{2.29}$$

When $\boldsymbol{Q}_{XX}$ is not invertible there are multiple solutions to (2.29). In this case the coefficient β is **not identified** as it does not have a unique value.

Minimization

The mean squared prediction error (2.19) is a function with vector argument of the form

$$f(x) = a - 2b'x + x'\boldsymbol{C}x$$

where $\boldsymbol{C} > 0$. For any function of this form, the unique minimizer is

$$x = \boldsymbol{C}^{-1}b. \tag{2.30}$$

To see that this is the unique minimizer we present two proofs. The first uses matrix calculus. From Appendix A.20

$$\frac{\partial}{\partial x}\left(b'x\right) = b \tag{2.31}$$

$$\frac{\partial}{\partial x}\left(x'\boldsymbol{C}x\right) = 2\boldsymbol{C}x \tag{2.32}$$

$$\frac{\partial^2}{\partial x \partial x'}\left(x'\boldsymbol{C}x\right) = 2\boldsymbol{C}. \tag{2.33}$$

Using (2.31) and (2.32), we find

$$\frac{\partial}{\partial x} f(x) = -2b + 2\boldsymbol{C}x.$$

The first-order condition for minimization sets this derivative equal to zero. Thus the solution satisfies $-2b + 2\boldsymbol{C}x = 0$. Solving for x we find (2.30). Using (2.33) we also find

$$\frac{\partial^2}{\partial x \partial x'} f(x) = 2\boldsymbol{C} > 0$$

which is the second-order condition for minimization. This shows that (2.30) is the unique minimizer of $f(x)$.

Our second proof is algebraic. Re-write $f(x)$ as

$$f(x) = \left(a - b'\boldsymbol{C}^{-1}b\right) + \left(x - \boldsymbol{C}^{-1}b\right)'\boldsymbol{C}\left(x - \boldsymbol{C}^{-1}b\right).$$

The first term does not depend on x so does not affect the minimizer. The second term is a quadratic form in a positive definite matrix. This means that for any non-zero α, $\alpha'\boldsymbol{C}\alpha > 0$. Thus for $x \neq \boldsymbol{C}^{-1}b$, the second-term is strictly positive, yet for $x = \boldsymbol{C}^{-1}b$ this term equals zero. It is therefore minimized at $x = \boldsymbol{C}^{-1}b$ as claimed.

2.19 Illustrations of Best Linear Predictor

We illustrate the best linear predictor (projection) using three log wage equations introduced in earlier sections.

For our first example, we consider a model with the two dummy variables for gender 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*.

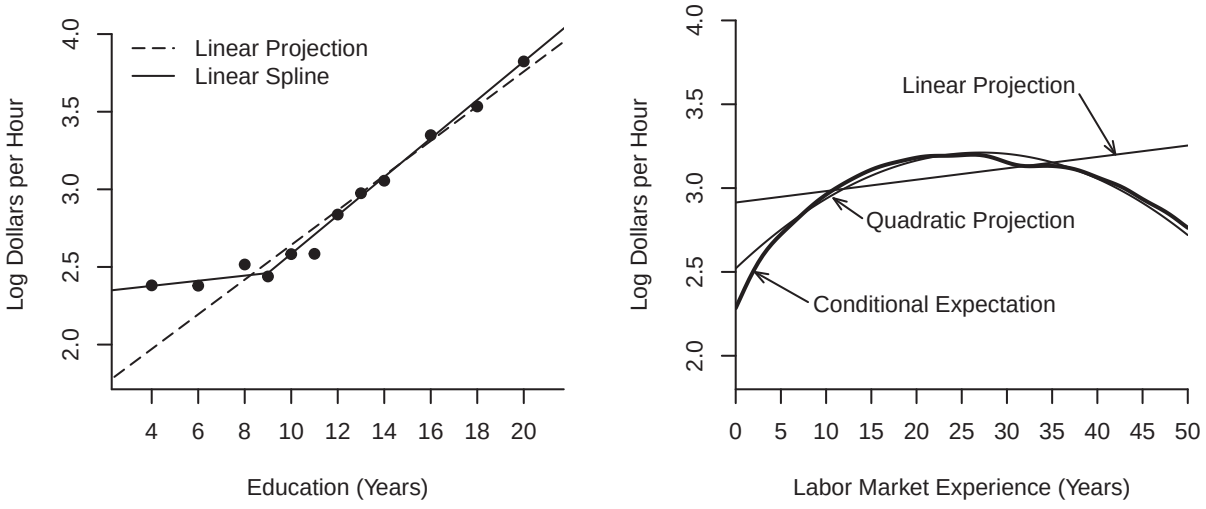
$$\mathbb{E}\left[\log(wage) \mid Black, female\right] = -0.20 Black - 0.24 female + 0.10 Black \times female + 3.06. \tag{2.34}$$

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*

$$\mathscr{P}\left[\log(wage) \mid Black, female\right] = -0.15 Black - 0.23 female + 3.06. \tag{2.35}$$

What is the difference? The full CEF (2.34) shows that the race gap is differentiated by gender: it is 20% for Black men (relative to non-Black men) and 10% for Black women (relative to non-Black women). The projection model (2.35) simplifies this analysis, calculating an average 15% wage gap for Black wage earners, ignoring the role of gender. Notice that this is despite the fact that *gender* is included in (2.35).



(a) Projections onto Education

(b) Projections onto Experience

Figure 2.6: Projections of Log Wage onto Education and Experience

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 2.3 and is repeated in Figure 2.6(a). 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

$$\mathscr{P}\left[\log(wage) \mid education\right] = 0.11 education + 1.5.$$

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:

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

This equation is displayed in Figure 2.6(a) using the solid line, and appears to fit much better. It indicates a 2% increase in mean wages for every year of education below 9, and a 12% increase in mean wages for every year of education above 9. It is still an approximation to the conditional mean but it appears to be fairly reasonable.

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 2.4 and is repeated as the solid line in Figure 2.6(b). Superimposed on the figure are two projections. The first (given by the dot-dashed line) is the linear projection on experience

$$\mathscr{P}\left[\log(wage) \mid experience\right] = 0.011 experience + 2.5$$

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

$$\mathscr{P}\left[\log(wage) \mid experience\right] = 0.046 experience - 0.0007 experience^2 + 2.3.$$

It is fairly clear from an examination of Figure 2.6(b) that the first linear projection is a poor approximation. It over-predicts wages for young and old workers, under-predicts for the rest, and 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** because the function is quadratic in *experience*.

2.20 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 $\boldsymbol{Q}_{YX} = \mathbb{E}\left[YX'\right]$ we can write σ^2 as

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left[\left(Y - X'\beta\right)^2\right] \\
&= \mathbb{E}\left[Y^2\right] - 2\mathbb{E}\left[YX'\right]\beta + \beta'\mathbb{E}\left[XX'\right]\beta \\
&= Q_{YY} - 2\boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} + \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&= Q_{YY} - \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&\stackrel{\text{def}}{=} Q_{YY\cdot X}.
\end{aligned} \tag{2.36}$$

One useful feature of this formula is that it shows that $Q_{YY\cdot X} = Q_{YY} - \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY}$ equals the variance of the error from the linear projection of Y on X.

2.21 Regression Coefficients

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

$$Y = X'\beta + \alpha + e \tag{2.37}$$

where α is the intercept and X does not contain a constant.

Taking expectations of this equation, we find

$$\mathbb{E}[Y] = \mathbb{E}\left[X'\beta\right] + \mathbb{E}[\alpha] + \mathbb{E}[e]$$

or $\mu_Y = \mu_X'\beta + \alpha$ where $\mu_Y = \mathbb{E}[Y]$ and $\mu_X = \mathbb{E}[X]$, since $\mathbb{E}[e] = 0$ from (2.28). (While X does not contain a constant, the equation does so (2.28) still applies.) Rearranging, we find $\alpha = \mu_Y - \mu_X'\beta$. Subtracting this equation from (2.37) we find

$$Y - \mu_Y = \left(X - \mu_X\right)'\beta + e, \tag{2.38}$$

a linear equation between the centered variables $Y-\mu_Y$ and $X-\mu_X$. (They are centered at their means so are mean-zero random variables.) Because $X-\mu_X$ is uncorrelated with e, (2.38) is also a linear projection. Thus by the formula for the linear projection model,

$$\begin{aligned}\beta &= \left(\mathbb{E}\left[\left(X-\mu_X\right)\left(X-\mu_X\right)'\right]\right)^{-1}\mathbb{E}\left[\left(X-\mu_X\right)\left(Y-\mu_Y\right)\right] \\ &= \operatorname{var}[X]^{-1}\operatorname{cov}(X,Y)\end{aligned}$$

a function only of the covariances[10] of X and Y.

Theorem 2.10 In the linear projection model $Y = X'\beta+\alpha+e$,

$$\alpha = \mu_Y - \mu_X'\beta \tag{2.39}$$

and

$$\beta = \operatorname{var}[X]^{-1}\operatorname{cov}(X,Y). \tag{2.40}$$

2.22 Regression Sub-Vectors

Let the regressors be partitioned as

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}. \tag{2.41}$$

We can write the projection of Y on X as

$$\begin{aligned} Y &= X'\beta + e \\ &= X_1'\beta_1 + X_2'\beta_2 + e \\ \mathbb{E}[Xe] &= 0. \end{aligned} \tag{2.42}$$

In this section we derive formulae for the sub-vectors β_1 and β_2.

Partition $\boldsymbol{Q}_{XX}$ conformably with X

$$\boldsymbol{Q}_{XX} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix} = \begin{bmatrix} \mathbb{E}\left[X_1X_1'\right] & \mathbb{E}\left[X_1X_2'\right] \\ \mathbb{E}\left[X_2X_1'\right] & \mathbb{E}\left[X_2X_2'\right] \end{bmatrix}$$

and similarly

$$\boldsymbol{Q}_{XY} = \begin{bmatrix} \boldsymbol{Q}_{1Y} \\ \boldsymbol{Q}_{2Y} \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X_1Y] \\ \mathbb{E}[X_2Y] \end{bmatrix}.$$

By the partitioned matrix inversion formula (A.3)

$$\boldsymbol{Q}_{XX}^{-1} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}^{-1} \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{Q}^{11} & \boldsymbol{Q}^{12} \\ \boldsymbol{Q}^{21} & \boldsymbol{Q}^{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}_{11\cdot2}^{-1} & -\boldsymbol{Q}_{11\cdot2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot1}^{-1} \end{bmatrix} \tag{2.43}$$

[10]The **covariance matrix** between vectors X and Z is $\operatorname{cov}(X,Z) = \mathbb{E}\left[(X-\mathbb{E}[X])(Z-\mathbb{E}[Z])'\right]$. The covariance matrix of the vector X is $\operatorname{var}[X] = \operatorname{cov}(X,X) = \mathbb{E}\left[(X-\mathbb{E}[X])(X-\mathbb{E}[X])'\right]$.

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

$$\begin{aligned}
\beta &= \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \\
&= \begin{bmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{Q}_{1Y} \\ \boldsymbol{Q}_{2Y} \end{bmatrix} \\
&= \begin{pmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\boldsymbol{Q}_{1y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2Y}\right) \\ \boldsymbol{Q}_{22\cdot 1}^{-1}\left(\boldsymbol{Q}_{2y} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{1Y}\right) \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1Y\cdot 2} \\ \boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{2Y\cdot 1} \end{pmatrix}.
\end{aligned}$$

We have shown that $\beta_1 = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1Y\cdot 2}$ and $\beta_2 = \boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{2Y\cdot 1}$.

2.23 Coefficient Decomposition

In the previous section we derived formulae for the coefficient sub-vectors β_1 and β_2. We now use these formulae to give a useful interpretation of the coefficients in terms of an iterated projection.

Take equation (2.42) for the case $\dim(X_1) = 1$ so that $\beta_1 \in \mathbb{R}$.

$$Y = X_1\beta_1 + X_2'\beta_2 + e. \tag{2.44}$$

Now consider the projection of X_1 on X_2 :

$$\begin{aligned}
X_1 &= X_2'\gamma_2 + u_1 \\
\mathbb{E}[X_2 u_1] &= 0.
\end{aligned}$$

From (2.22) and (2.36), $\gamma_2 = \boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and $\mathbb{E}\left[u_1^2\right] = \boldsymbol{Q}_{11\cdot 2} = \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$. We can also calculate that

$$\mathbb{E}[u_1 Y] = \mathbb{E}\left[\left(X_1 - \gamma_2' X_2\right) Y\right] = \mathbb{E}[X_1 Y] - \gamma_2'\mathbb{E}[X_2 Y] = \boldsymbol{Q}_{1Y} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{2Y} = \boldsymbol{Q}_{1Y\cdot 2}.$$

We have found that

$$\beta_1 = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{1Y\cdot 2} = \frac{\mathbb{E}[u_1 Y]}{\mathbb{E}\left[u_1^2\right]}$$

the coefficient from the simple regression of Y on u_1.

What this means is that in the multivariate projection equation (2.44), 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 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.

2.24 Omitted Variable Bias

Again, let the regressors be partitioned as in (2.41). Consider the projection of Y on X_1 only. Perhaps this is done because the variables X_2 are not observed. This is the equation

$$\begin{aligned} Y &= X_1'\gamma_1 + u \\ \mathbb{E}[X_1 u] &= 0. \end{aligned} \tag{2.45}$$

Notice that we have written the coefficient as γ_1 rather than β_1 and the error as u rather than e. This is because (2.45) is different than (2.42). Goldberger (1991) introduced the catchy labels **long regression** for (2.42) and **short regression** for (2.45) to emphasize the distinction.

Typically, $\beta_1 \neq \gamma_1$, except in special cases. To see this, we calculate

$$\begin{aligned} \gamma_1 &= \left(\mathbb{E}\left[X_1 X_1'\right]\right)^{-1} \mathbb{E}[X_1 Y] \\ &= \left(\mathbb{E}\left[X_1 X_1'\right]\right)^{-1} \mathbb{E}\left[X_1\left(X_1'\beta_1 + X_2'\beta_2 + e\right)\right] \\ &= \beta_1 + \left(\mathbb{E}\left[X_1 X_1'\right]\right)^{-1} \mathbb{E}\left[X_1 X_2'\right]\beta_2 \\ &= \beta_1 + \Gamma_{12}\beta_2 \end{aligned}$$

where $\Gamma_{12} = \boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$ is the coefficient matrix from a projection of X_2 on X_1 where we use the notation from Section 2.22.

Observe that $\gamma_1 = \beta_1 + \Gamma_{12}\beta_2 \neq \beta_1$ unless $\Gamma_{12} = 0$ or $\beta_2 = 0$. Thus the short and long regressions have different coefficients. They are the same only under one of two conditions. First, if the projection of X_2 on X_1 yields a set of zero coefficients (they are uncorrelated), or second, if the coefficient on X_2 in (2.42) is zero. The difference $\Gamma_{12}\beta_2$ between γ_1 and β_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 $\Gamma_{12} > 0$. It also seems reasonable to suppose that conditional on education, individuals with higher intelligence will earn higher wages on average, so that $\beta_2 > 0$. This implies that $\Gamma_{12}\beta_2 > 0$ and $\gamma_1 = \beta_1 + \Gamma_{12}\beta_2 > \beta_1$. Therefore, it seems reasonable to expect that in a regression of wages on education with intelligence omitted (as the latter is not measured), the coefficient on education is higher than in a regression where intelligence is included. In other words, in this context the omitted variable biases the regression coefficient upwards. It is possible, for example, that $\beta_1 = 0$ so that education has no direct effect on wages yet $\gamma_1 = \Gamma_{12}\beta_2 > 0$ meaning that the regression coefficient on education alone is positive, but is a consequence of the unmodeled correlation between education and intellectual ability.

Unfortunately, the above simple characterization of omitted variable bias does not immediately carry over to more complicated settings, as discovered by Luca, Magnus, and Peracchi (2018). For example, suppose we compare three nested projections

$$\begin{aligned} Y &= X_1'\gamma_1 + u_1 \\ Y &= X_1'\delta_1 + X_2'\delta_2 + u_2 \\ Y &= X_1'\beta_1 + X_2'\beta_2 + X_3'\beta_3 + e. \end{aligned}$$

We can call them short, medium, and long regressions. Suppose that the parameter of interest is β_1 in the long regression. We are interested in the consequences of omitting X_3 when estimating the medium regression, and of omitting both X_2 and X_3 when estimating the short regression. In particular we are interested in the question: Is it better to estimate the short or medium regression, given that both omit X_3? Intuition suggests that the medium regression should be "less biased" but it is worth investigating in greater detail. By similar calculations to those above, we find that

$$\gamma_1 = \beta_1 + \Gamma_{12}\beta_2 + \Gamma_{13}\beta_3$$
$$\delta_1 = \beta_1 + \Gamma_{13\cdot 2}\beta_3$$

where $\Gamma_{13\cdot 2} = \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{13\cdot 2}$ using the notation from Section 2.22.

We see that the bias in the short regression coefficient is $\Gamma_{12}\beta_2 + \Gamma_{13}\beta_3$ which depends on both β_2 and β_3, while that for the medium regression coefficient is $\Gamma_{13\cdot 2}\beta_3$ which only depends on β_3. So the bias for the medium regression is less complicated and intuitively seems more likely to be smaller than that of the short regression. However it is impossible to strictly rank the two. It is quite possible that γ_1 is less biased than δ_1. Thus as a general rule it is unknown if estimation of the medium regression will be less biased than estimation of the short regression.

2.25 Best Linear Approximation

There are alternative ways we could construct a linear approximation $X'\beta$ to the conditional expectation $m(X)$. 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 $X'\beta$ to $m(X)$ as the expected squared difference between $X'\beta$ and the conditional expectation $m(X)$

$$d(\beta) = \mathbb{E}\left[\left(m(X) - X'\beta\right)^2\right].$$

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

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

where $f_X(x)$ is the marginal density of X.

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

$$\beta = \operatorname*{argmin}_{b\in\mathbb{R}^k} d(b). \tag{2.46}$$

Similar to the best linear predictor we are measuring accuracy by expected squared error. The difference is that the best linear predictor (2.18) selects β to minimize the expected squared prediction error, while the best linear approximation (2.46) selects β 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

$$\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[Xm(X)\right] \tag{2.47}$$
$$= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right] \tag{2.48}$$

(see Exercise 2.19). Thus (2.46) equals (2.18). We conclude that the definition (2.46) can be viewed as an alternative motivation for the linear projection coefficient.

2.26 Regression to the Mean

The term **regression** originated in an influential paper by Francis Galton (1886) where he examined the joint distribution of the stature (height) of parents and children. Effectively, he was estimating the conditional expectation of children's height given their parent's height. Galton discovered that this conditional expectation 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

$$Y = X\beta + \alpha + e \tag{2.49}$$

where Y equals the height of the child and X equals the height of the parent. Assume that Y and X have the same expectation so that $\mu_Y = \mu_X = \mu$. Then from (2.39) $\alpha = \left(1-\beta\right)\mu$ so we can write the linear projection (2.49) as

$$\mathscr{P}(Y \mid X) = \left(1-\beta\right)\mu + X\beta.$$

This shows that the projected height of the child is a weighted average of the population expectation μ and the parent's height X with weights β and $1-\beta$. 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 (2.40)

$$\beta = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}[X]} = \mathrm{corr}(X,Y).$$

By the Cauchy-Schwarz inequality (B.32), $-1 \le \mathrm{corr}(X,Y) \le 1$, 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.

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 (2.49). Since X and e are uncorrelated, it follows that

$$\mathrm{var}[Y] = \beta^2 \mathrm{var}[X] + \mathrm{var}[e].$$

Then $\mathrm{var}[Y] < \mathrm{var}[X]$ if and only if

$$\beta^2 < 1 - \frac{\mathrm{var}[e]}{\mathrm{var}[X]}$$

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

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

$$X = Y\beta^* + \alpha^* + e^*. \tag{2.50}$$

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

$$\beta^* = \operatorname{corr}(X, Y) = \beta$$

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

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! [This equality is not particularly imporant; it is an artifact of the assumption that X and Y have the same variances.]

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 (2.49), divide through by β and rewrite to find the equation

$$X = Y\frac{1}{\beta} - \frac{\alpha}{\beta} - \frac{1}{\beta}e \tag{2.51}$$

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 (2.51) is perfectly valid because it is a simple manipulation of the valid equation (2.49). The trouble is that (2.51) is neither a CEF nor a linear projection. Inverting a projection (or CEF) does not yield a projection (or CEF). Instead, (2.50) is a valid projection, not (2.51).

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

2.28 Limitations of the Best Linear Projection

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

From Theorem 2.4.4 we know that the CEF error has the property $\mathbb{E}[Xe] = 0$. Thus a linear CEF is the best linear projection. However, the converse is not true as the projection error does not necessarily satisfy $\mathbb{E}[e \mid X] = 0$. 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 $Y = \beta X + \alpha + e$. Since $X \sim \mathrm{N}(0,1)$ then X and X^2 are uncorrelated and the linear projection takes the form $\mathscr{P}[Y \mid X] = X + 1$. 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.

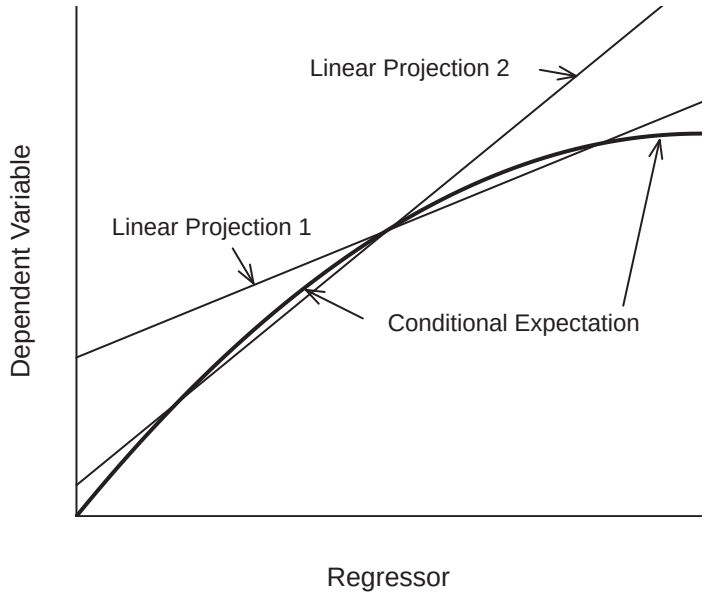


Figure 2.7: Conditional Expectation and Two Linear Projections

Another defect of linear projection is that it is sensitive to the marginal distribution of the regressors when the conditional mean is nonlinear. We illustrate the issue in Figure 2.7 for a constructed[11] joint distribution of Y and X. The thick line is the nonlinear CEF of Y given X. The data are divided in two groups – Group 1 and Group 2 – which have different marginal distributions for the regressor X, and Group 1 has a lower mean value of X than Group 2. The separate linear projections of Y on X for these two groups are displayed in the figure by the thin 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 expectation function. The apparent difference is a by-product of linear approximations to a nonlinear expectation combined with different marginal distributions for the conditioning variables.

[11] The X in Group 1 are $\mathrm{N}(2,1)$, those in Group 2 are $\mathrm{N}(4,1)$, and the conditional distribution of Y given X is $\mathrm{N}(m(X),1)$ where $m(x) = 2x - x^2/6$. The functions are plotted over $0 \le x \le 6$.

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 $Y = X'\eta$ where the individual-specific coefficient η is random and independent of X. For example, if X is years of schooling and Y is log wages, then η is the individual-specific returns to schooling. If a person obtains an extra year of schooling, η 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 η) 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 expectation due to a change in the regressors, $\beta = \nabla m(X)$. 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 $\eta = \nabla\left(X'\eta\right)$ 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 $\beta = \mathbb{E}\left[\eta\right]$ and $\Sigma = \operatorname{var}\left[\eta\right]$ denote the mean and covariance matrix of η and then decompose the random coefficient as $\eta = \beta + u$ where u is distributed independently of X with mean zero and covariance matrix Σ. Then we can write

$$\mathbb{E}\left[Y \mid X\right] = X'\mathbb{E}\left[\eta \mid X\right] = X'\mathbb{E}\left[\eta\right] = X'\beta$$

so the CEF is linear in X, and the coefficient β equals the expectation of the random coefficient η.

We can thus write the equation as a linear CEF $Y = X'\beta + e$ where $e = X'u$ and $u = \eta - \beta$. The error is conditionally mean zero: $\mathbb{E}\left[e \mid X\right] = 0$. Furthermore

$$\operatorname{var}\left[e \mid X\right] = X'\operatorname{var}\left[\eta\right]X = X'\Sigma X$$

so the error is conditionally heteroskedastic with its variance a quadratic function of X.

Theorem 2.11 In the linear random coefficient model $Y = X'\eta$ with η independent of X, $\mathbb{E}\left\|X\right\|^2 < \infty$, and $\mathbb{E}\left\|\eta\right\|^2 < \infty$, then

$$\mathbb{E}\left[Y \mid X\right] = X'\beta$$
$$\operatorname{var}\left[Y \mid X\right] = X'\Sigma X$$

where $\beta = \mathbb{E}\left[\eta\right]$ and $\Sigma = \operatorname{var}\left[\eta\right]$.

2.30 Causal Effects

So far we have avoided the concept of causality, yet often the underlying goal of an econometric analysis is to measure 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 due to a change in

an input. We are not just interested in the conditional expectation or linear projection, we would like to know the actual change.

Two inherent barriers are: (1) the causal effect is typically specific to an individual; and (2) the causal effect is typically 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 are 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 concrete suppose that there are two individuals, Jennifer and George, and both have the possibility of being high-school graduates or college graduates, and 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.

Rubin (1974) developed the **potential outcomes** framework (also known as the **Rubin causal model**) to clarify the issues. Let Y be a scalar outcome (for example, wages) and D be a binary **treatment** (for example, college attendence). The specification of treatment as binary is not essential but simplifies the notation. A flexible model describing the impact of the treatment on the outcome is

$$Y = h(D, U) \tag{2.52}$$

where U is an $\ell \times 1$ unobserved random factor and h is a functional relationship. It is also common to use the simplified notation $Y(0) = h(0, U)$ and $Y(1) = h(1, U)$ for the potential outcomes associated with non-treatment and treatment, respectively. The notation implicitly holds U fixed. The potential outcomes are specific to each individual as they depend on U. For example, if Y is an individual's wage, the unobservables U could include characteristics such as the individual's abilities, skills, work ethic, interpersonal connections, and preferences, all of which potentially influence their wage. In our example these factors are summarized by the labels "Jennifer" and "George".

Rubin described the effect as **causal** when we vary D while holding U constant. In our example this means changing an individual's education while holding constant their other attributes.

Definition 2.6 In the model (2.52) the **causal effect** of D on Y is

$$C(U) = Y(1) - Y(0) = h(1, U) - h(0, U), \tag{2.53}$$

the change in Y due to treatment while holding U constant.

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).

The causal effect of treatment $C(U)$ defined in (2.53) is heterogeneous and random as the potential outcomes $Y(0)$ and $Y(1)$ vary across individuals. Also, we do not observe both $Y(0)$ and $Y(1)$ for a given individual, but rather only the realized value

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

Table 2.3: Example Distribution

	$8	$10	$12	$20	Mean
High-School Graduate	10	6	0	0	$8.75
College Graduate	0	0	6	10	$17.00
Difference					$8.25

Consequently the causal effect $C(U)$ is unobserved.

Rubin's goal was to learn features of the distribution of $C(U)$ including its expected value which he called the average causal effect. He defined it as follows.

Definition 2.7 In the model (2.52) the **average causal effect** of D on Y is

$$\text{ACE} = \mathbb{E}[C(U)] = \int_{\mathbb{R}^\ell} C(u) f(u) du$$

where $f(u)$ is the density of U.

The ACE is the population average of the causal effect. Extending our Jennifer & George example, suppose that half of the population are like Jennifer and the other half are like George. Then the average causal effect of college on wages is $(10+4)/2 = \$7$ an hour.

To estimate the ACE a reasonable starting place is to compare average Y for treated and untreated individuals. In our example this is the difference between the average wage among college graduates and high school graduates. This is the same as the coefficient in a regression of the outcome Y on the treatment D. Does this equal the ACE?

The answer depends on the relationship between treatment D and the unobserved component U. If D is randomly assigned as in an experiment then D and U are independent and the regression coefficient equals the ACE. However, if D and U are dependent then the regression coefficient and ACE are different. To see this, observe that the difference between the average outcomes of the treated and untreated populations are

$$\mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0] = \int_{\mathbb{R}^\ell} h(1,u) f(u \mid D=1)\, du - \int_{\mathbb{R}^\ell} h(1,u) f(u \mid D=0)\, du$$

where $f(u \mid D)$ is the conditional density of U given D. If U is independent of D then $f(u \mid D) = f(u)$ and the above expression equals $\int_{\mathbb{R}^\ell} (h(1,u) - h(0,u))\, f(u)\, du = \text{ACE}$. However, if U and D are dependent this equality fails.

To illustrate, let's return to our example of Jennifer and George. Suppose that all high school students take an aptitude test. If a student gets a high (H) score they go to college with probability 3/4, and if a student gets a low (L) score they go to college with probability 1/4. Suppose further that Jennifer gets an aptitude score of H with probability 3/4, while George gets a score of H with probability 1/4. Given this situation, 62.5% of Jennifer's will go to college[12] while 37.5% of George's will go to college[13].

An econometrician who randomly samples 32 individuals and collects data on educational attainment and wages will find the wage distribution displayed in Table 2.3.

[12] $\mathbb{P}[college \mid Jennifer] = \mathbb{P}[college \mid H]\,\mathbb{P}[H \mid Jennifer] + \mathbb{P}[college \mid L]\,\mathbb{P}[L \mid Jennifer] = (3/4)^2 + (1/4)^2$.

[13] $\mathbb{P}[college \mid George] = \mathbb{P}[college \mid H]\,\mathbb{P}[H \mid George] + \mathbb{P}[college \mid L]\,\mathbb{P}[L \mid George] = (3/4)(1/4) + (1/4)(3/4)$.

Our econometrician finds that the average wage among high school graduates is $8.75 while the average wage among college graduates is $17.00. The difference of $8.25 is the econometrician's regression coefficient for the effect of college on wages. But $8.25 overstates the true ACE of $7. The reason is that college attendence is determined by an aptitude test which is correlated with an individual's causal effect. Jennifer has both a high causal effect and is more likely to attend college, so the observed difference in wages overstates the causal effect of college.

To visualize Table 2.3 examine Figure 2.8. The four points are the four education/wage pairs from the table, with the size of the points calibrated to the wage distribution. The two lines are the econometrician's regression line and the average causal effect. The Jennifer's in the population correspond to the points above the two lines, the George's in the population correspond to the points below the two lines. Because most Jennifer's go to College, and most George's do not, the regression line is tilted away from the average causal effect towards the two large points.

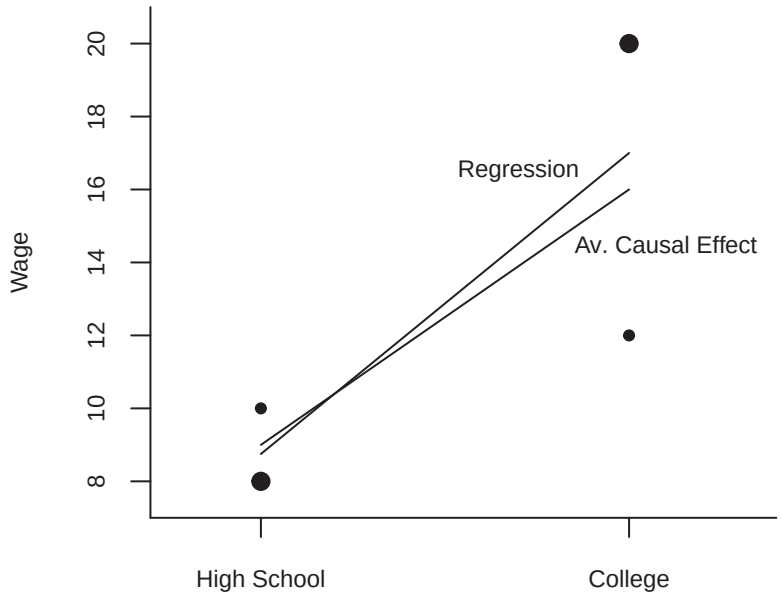


Figure 2.8: Average Causal Effect vs Regression

Our first lesson from this analysis is that we need to be cautious about interpreting regression coefficients as causal effects. Unless the regressors (e.g. education attainment) can be interpreted as randomly assigned it is inappropriate to interpret the regression coefficients causally.

Our second lesson will be that a causal interpretation can be obtained if we condition on a sufficiently rich set of covariates. We now explore this issue.

Suppose that the observables include a set of covariates X in addition to the outcome Y and treatment D. We extend the potential outcomes model (2.52) to include X:

$$Y = h(D, X, U). \tag{2.54}$$

We also extend the definition of a causal effect to allow conditioning on X.

Definition 2.8 In the model (2.54) the **causal effect** of D on Y is

$$C(X,U) = h(1,X,U) - h(0,X,U),$$

the change in Y due to treatment holding X and U constant.
The **conditional average causal effect** of D on Y conditional on $X = x$ is

$$\mathrm{ACE}(x) = \mathbb{E}[C(X,U) \mid X = x] = \int_{\mathbb{R}^\ell} C(x,u) f(u \mid x) du$$

where $f(u \mid x)$ is the conditional density of U given X.
The **unconditional average causal effect** of D on Y is

$$\mathrm{ACE} = \mathbb{E}[C(X,U)] = \int \mathrm{ACE}(x) f(x) dx$$

where $f(x)$ is the density of X.

The conditional average causal effect $\mathrm{ACE}(x)$ is the ACE for the sub-population with characteristics $X = x$. Given observations on (Y, D, X) we want to measure the causal effect of D on Y, and are interested if this can be obtained by a regression of Y on (D, X). We would like to interpret the coefficient on D as a causal effect. Is this appropriate?

Our previous analysis showed that a causal interpretation obtains when U is independent of the regressors. While this is sufficient it is stronger than necessary. Instead, the following is sufficient.

Definition 2.9 Conditional Independence Assumption (CIA). Conditional on X, the random variables D and U are statistically independent.

The CIA implies that the conditional density of U given (D, X) only depends on X, thus $f(u \mid D, X) = f(u \mid X)$. This implies that the regression of Y on (D, X) equals

$$\begin{aligned} m(d,x) &= \mathbb{E}[Y \mid D = d, X = x] \\ &= \mathbb{E}[h(d,x,U) \mid D = d, X = x] \\ &= \int h(d,x,u) f(u \mid x) du. \end{aligned}$$

Under the CIA the treatment effect measured by the regression is

$$\begin{aligned} \nabla m(d,x) &= m(1,x) - m(0,x) \\ &= \int h(1,x,u) f(u \mid x) du - \int h(0,x,u) f(u \mid x) du \\ &= \int C(x,u) f(u \mid x) du \\ &= \mathrm{ACE}(x). \end{aligned} \tag{2.55}$$

This is the conditional ACE. Thus under the CIA the regression coefficient equals the ACE.

We deduce that the regression of Y on (D, X) reveals the causal impact of treatment when the CIA holds. This means that regression analysis can be interpreted causally when we can make the case that the regressors X are sufficient to control for factors which are correlated with treatment.

Theorem 2.12 In the structural model (2.54), the Conditional Independence Assumption implies $\nabla m(d, x) = \text{ACE}(x)$, that the regression derivative with respect to treatment equals the conditional ACE.

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 conditional causal effect. This means the CEF has causal economic meaning, giving strong justification to estimation of the CEF.

It is important to understand the critical role of the CIA. If CIA fails then the equality (2.55) of the regression derivative and the ACE fails. The CIA states that conditional on X the variables U and D are independent. This means that treatment D is not affected by the unobserved individual factors U and is effectively random. It is a strong assumption. In the wage/education example it means that education is not selected by individuals based on their unobserved characteristics.

However, it is also helpful to understand that the CIA is weaker than full independence of U from the regressors (D, X). What is required is only that U and D are independent after conditioning on X. If X is sufficiently rich this may not be restrictive.

Returning to our example, we require a variable X which breaks the dependence between D and U. In our example, this variable is the aptitude test score, because the decision to attend college was based on the test score. It follows that educational attainment and type are independent once we condition on the test score.

To see this, observe that if a student's test score is H the probability they go to college ($D = 1$) is $3/4$ for both Jennifers and Georges. Similarly, if their test score is L the probability they go to college is $1/4$ for both types. This means that college attendence is independent of type, conditional on the aptitude test score.

The conditional ACE depends on the test score. Among students who receive a high test score, $3/4$ are Jennifers and $1/4$ are Georges. Thus the conditional ACE for students with a score of H is $(3/4) \times 10 + (1/4) \times 4 = \$8.50$. Among students who receive a low test score, $1/4$ are Jennifers and $3/4$ are Georges. Thus the ACE for students with a score of L is $(1/4) \times 10 + (3/4) \times 4 = \$5.50$. The unconditional ACE is the average, $\text{ACE} = (8.50 + 5.50)/2 = \$7$, because 50% of students each receive scores of H and L.

Theorem 2.12 shows that the conditional ACE is revealed by a regression which includes 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 wage distribution in Table 2.4.

Define a dummy *highscore* to indicate students who received a high test score. The regression of wages on college attendance and test scores with their interaction is

$$\mathbb{E}\left[\textit{wage} \mid \textit{college}, \textit{highscore}\right] = 1.00\textit{highscore} + 5.50\textit{college} + 3.00\textit{highscore} \times \textit{college} + 8.50. \tag{2.56}$$

The coefficient on *college*, \$5.50, is the regression derivative of college attendance for those with low test scores, and the sum of this coefficient with the interaction coefficient \$3.00 equals \$8.50 which is the

Table 2.4: Example Distribution 2

	$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

regression derivative for college attendance for those with high test scores. $5.50 and $8.50 equal the conditional causal effects as calculated above.

This shows that from the regression (2.56) an econometrician will find that the effect of college on wages is $8.50 for those with high test scores and $5.50 for those with low test scores with an average effect of $7 (because 50% of students receive high and low test scores). This is the true average causal effect of college on wages. Thus the regression coefficient on *college* in (2.56) can be interpreted causally, while a regression omitting the aptitude test score does not reveal the causal effect of education.

To summarize our findings, we have shown how it is possible that a simple regression will give a false measurement of a causal effect, but a more careful regression can reveal the true causal effect. The key is to condition on a suitably rich set of covariates such that the remaining unobserved factors affecting the outcome are independent of the treatment variable.

2.31 Existence and Uniqueness of the Conditional Expectation*

In Sections 2.3 and 2.6 we defined the conditional expectation when the conditioning variables X are discrete and when the variables (Y, 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, X) for which Y has a finite mean.

Theorem 2.13 Existence of the Conditional Expectation
If $\mathbb{E}|Y| < \infty$ then there exists a function $m(x)$ such that for all sets $\mathscr{X}$ for which $\mathbb{P}[X \in \mathscr{X}]$ is defined,

$$\mathbb{E}[\mathbb{1}\{X \in \mathscr{X}\} Y] = \mathbb{E}[\mathbb{1}\{X \in \mathscr{X}\} m(X)]. \tag{2.57}$$

The function $m(X)$ is almost everywhere unique, in the sense that if $h(x)$ satisfies (2.57), then there is a set S such that $\mathbb{P}[S] = 1$ and $m(x) = h(x)$ for $x \in S$. The function $m(x)$ is called the **conditional expectation** and is written $m(x) = \mathbb{E}[Y \mid X = x]$.

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

The conditional expectation $m(x)$ defined by (2.57) specializes to (2.6) when (Y, X) have a joint density. The usefulness of definition (2.57) is that Theorem 2.13 shows that the conditional expectation $m(X)$

exists for all finite-mean distributions. This definition allows Y to be discrete or continuous, for X to be scalar or vector-valued, and for the components of X to be discrete or continuously distributed.

You may have noticed that Theorem 2.13 applies only to sets $\mathscr{X}$ for which $\mathbb{P}[X \in \mathscr{X}]$ is defined. This is a technical issue – measurability – which we largely side-step in this textbook. Formal probability theory only applies to sets which are measurable – for which probabilities are defined – as it turns out that not all sets satisfy measurability. This is not a practical concern for applications, so we defer such distinctions for formal theoretical treatments.

2.32 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 expectation, 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 $\mathscr{F}$ be a collection of distributions F. Let θ be a parameter of interest (for example, the expectation $\mathbb{E}[Y]$).

Definition 2.10 A parameter $\theta \in \mathbb{R}$ is identified on $\mathscr{F}$ if for all $F \in \mathscr{F}$, 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 $\mathscr{F}$. The restriction to the set $\mathscr{F}$ is important. Most parameters are identified only on a strict subset of the space of all distributions.

Take, for example, the expectation $\mu = \mathbb{E}[Y]$. It is uniquely determined if $\mathbb{E}|Y| < \infty$, so μ is identified for the set $\mathscr{F} = \{F : \mathbb{E}|Y| < \infty\}$.

Next, consider the conditional expectation. Theorem 2.13 demonstrates that $\mathbb{E}|Y| < \infty$ is a sufficient condition for identification.

Theorem 2.14 Identification of the Conditional Expectation
If $\mathbb{E}|Y| < \infty$, the conditional expectation $m(x) = \mathbb{E}[Y \mid X = x]$ 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

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

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

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

We are interested in features of the distribution F not the censored distribution F^*. For example, we are interested in the expected 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 $\mathbb{P}[Y \geq \tau] = 0$. Thus the expectation 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 $\mathscr{F}$ 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 $F \in \mathscr{F}$. 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 $\alpha \leq \mathbb{P}[Y \leq \tau]$. 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.

2.33 Technical Proofs*

Proof of Theorem 2.1 For convenience, assume that the variables have a joint density $f(y, x)$. Since $\mathbb{E}[Y \mid X]$ is a function of the random vector X only, to calculate its expectation we integrate with respect to the density $f_X(x)$ of X, that is

$$\mathbb{E}[\mathbb{E}[Y \mid X]] = \int_{\mathbb{R}^k} \mathbb{E}[Y \mid X] f_X(x)\, dx.$$

Substituting in (2.6) and noting that $f_{Y|X}(y \mid x) f_X(x) = f(y, x)$, we find that the above expression equals

$$\int_{\mathbb{R}^k} \left(\int_{\mathbb{R}} y f_{Y|X}(y \mid x)\, dy \right) f_X(x)\, dx = \int_{\mathbb{R}^k} \int_{\mathbb{R}} y f(y, x)\, dy dx = \mathbb{E}[Y]$$

the unconditional expectation of Y. ■

Proof of Theorem 2.2 Again assume that the variables have a joint density. It is useful to observe that

$$f(y \mid x_1, x_2) f(x_2 \mid x_1) = \frac{f(y, x_1, x_2)}{f(x_1, x_2)} \frac{f(x_1, x_2)}{f(x_1)} = f(y, x_2 \mid x_1), \tag{2.58}$$

the density of (Y, X_2) given 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

$$\mathbb{E}[Y \mid X_1 = x_1, X_2 = x_2] = \int_{\mathbb{R}} y f(y \mid x_1, x_2)\, dy. \tag{2.59}$$

Integrating (2.59) with respect to the conditional density of X_2 given X_1, and applying (2.58) we find that

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[Y \mid X_1, X_2\right] \mid X_1 = x_1\right] &= \int_{\mathbb{R}^{k_2}} \mathbb{E}\left[Y \mid X_1 = x_1, X_2 = x_2\right] f\left(x_2 \mid x_1\right) dx_2 \\
&= \int_{\mathbb{R}^{k_2}} \left(\int_{\mathbb{R}} y f\left(y \mid x_1, x_2\right) dy \right) f\left(x_2 \mid x_1\right) dx_2 \\
&= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f\left(y \mid x_1, x_2\right) f\left(x_2 \mid x_1\right) dy dx_2 \\
&= \int_{\mathbb{R}^{k_2}} \int_{\mathbb{R}} y f\left(y, x_2 \mid x_1\right) dy dx_2 \\
&= \mathbb{E}\left[Y \mid X_1 = x_1\right].
\end{aligned}$$

This implies $\mathbb{E}\left[\mathbb{E}\left[Y \mid X_1, X_2\right] \mid X_1\right] = \mathbb{E}\left[Y \mid X_1\right]$ as stated. ■

Proof of Theorem 2.3

$$\mathbb{E}\left[g\left(X\right) Y \mid X = x\right] = \int_{\mathbb{R}} g\left(x\right) y f_{Y|X}\left(y \mid x\right) dy = g\left(x\right) \int_{\mathbb{R}} y f_{Y|X}\left(y \mid x\right) dy = g\left(x\right) \mathbb{E}\left[Y \mid X = x\right]$$

This implies $\mathbb{E}\left[g\left(X\right) Y \mid X\right] = g\left(X\right) \mathbb{E}\left[Y \mid X\right]$ which is (2.7). Equation (2.8) follows by applying the simple law of iterated expectations (Theorem 2.1) to (2.7). ■

Proof of Theorem 2.4 Applying Minkowski's inequality (B.34) to $e = Y - m(X)$,

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

where the two parts on the right-hand-side are finite because $\mathbb{E}\left|Y\right|^r < \infty$ by assumption and $\mathbb{E}\left|m(X)\right|^r < \infty$ by the conditional expectation inequality (B.29). 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 2.6 The assumption that $\mathbb{E}\left[Y^2\right] < \infty$ implies that all the conditional expectations below exist.

Using the law of iterated expectations (Theorem 2.2) $\mathbb{E}\left[Y \mid X_1\right] = \mathbb{E}\left(\mathbb{E}\left[Y \mid X_1, X_2\right] \mid X_1\right)$ and the conditional Jensen's inequality (B.28),

$$\left(\mathbb{E}\left[Y \mid X_1\right]\right)^2 = \left(\mathbb{E}\left(\mathbb{E}\left[Y \mid X_1, X_2\right] \mid X_1\right)\right)^2 \leq \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid X_1, X_2\right]\right)^2 \mid X_1\right].$$

Taking unconditional expectations, this implies

$$\mathbb{E}\left[\left(\mathbb{E}\left[Y \mid X_1\right]\right)^2\right] \leq \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid X_1, X_2\right]\right)^2\right].$$

Similarly,

$$\left(\mathbb{E}\left[Y\right]\right)^2 \leq \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid X_1\right]\right)^2\right] \leq \mathbb{E}\left[\left(\mathbb{E}\left[Y \mid X_1, X_2\right]\right)^2\right]. \tag{2.60}$$

The variables Y, $\mathbb{E}\left[Y \mid X_1\right]$, and $\mathbb{E}\left[Y \mid X_1, X_2\right]$ all have the same expectation $\mathbb{E}\left[Y\right]$, so the inequality (2.60) implies that the variances are ranked monotonically:

$$0 \leq \operatorname{var}\left(\mathbb{E}\left[Y \mid X_1\right]\right) \leq \operatorname{var}\left(\mathbb{E}\left[Y \mid X_1, X_2\right]\right). \tag{2.61}$$

Define $e = Y - \mathbb{E}\left[Y \mid X\right]$ and $u = \mathbb{E}\left[Y \mid X\right] - \mu$ so that we have the decomposition $Y - \mu = e + u$. Notice $\mathbb{E}\left[e \mid X\right] = 0$ and u is a function of X. Thus by the conditioning theorem (Theorem 2.3), $\mathbb{E}\left[eu\right] = 0$ so e and u are uncorrelated. It follows that

$$\operatorname{var}\left[Y\right] = \operatorname{var}\left[e\right] + \operatorname{var}\left[u\right] = \operatorname{var}\left[Y - \mathbb{E}\left[Y \mid X\right]\right] + \operatorname{var}\left[\mathbb{E}\left[Y \mid X\right]\right]. \tag{2.62}$$

The monotonicity of the variances of the conditional expectation (2.61) applied to the variance decomposition (2.62) implies the reverse monotonicity of the variances of the differences, completing the proof. ■

Proof of Theorem 2.9 For part 1, by the expectation inequality (B.30), (A.17) and Assumption 2.1,

$$\left\|\mathbb{E}\left[XX'\right]\right\| \leq \mathbb{E}\left\|XX'\right\| = \mathbb{E}\left\|X\right\|^2 < \infty.$$

Similarly, using the expectation inequality (B.30), the Cauchy-Schwarz inequality (B.32), and Assumption 2.1,

$$\left\|\mathbb{E}\left[XY\right]\right\| \leq \mathbb{E}\left\|XY\right\| \leq \left(\mathbb{E}\left\|X\right\|^2\right)^{1/2}\left(\mathbb{E}\left[Y^2\right]\right)^{1/2} < \infty.$$

Thus the moments $\mathbb{E}\left[XY\right]$ and $\mathbb{E}\left[XX'\right]$ are finite and well defined.

For part 2, the coefficient $\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right]$ is well defined because $\left(\mathbb{E}\left[XX'\right]\right)^{-1}$ exists under Assumption 2.1.

Part 3 follows from Definition 2.5 and part 2.

For part 4, first note that

$$\begin{aligned}
\mathbb{E}\left[e^2\right] &= \mathbb{E}\left[\left(Y - X'\beta\right)^2\right] \\
&= \mathbb{E}\left[Y^2\right] - 2\mathbb{E}\left[YX'\right]\beta + \beta'\mathbb{E}\left[XX'\right]\beta \\
&= \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[YX'\right]\left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right] \\
&\leq \mathbb{E}\left[Y^2\right] < \infty.
\end{aligned}$$

The first inequality holds because $\mathbb{E}\left[YX'\right]\left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right]$ is a quadratic form and therefore necessarily non-negative. Second, by the expectation inequality (B.30), the Cauchy-Schwarz inequality (B.32), and Assumption 2.1,

$$\left\|\mathbb{E}\left[Xe\right]\right\| \leq \mathbb{E}\left\|Xe\right\| = \left(\mathbb{E}\left\|X\right\|^2\right)^{1/2}\left(\mathbb{E}\left[e^2\right]\right)^{1/2} < \infty.$$

It follows that the expectation $\mathbb{E}\left[Xe\right]$ is finite, and is zero by the calculation (2.26).

For part 6, applying Minkowski's inequality (B.34) to $e = Y - X'\beta$,

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

the final inequality by assumption. ■

2.34 Exercises

Exercise 2.1 Find $\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[Y \mid X_1, X_2, X_3\right] \mid X_1, X_2\right] \mid X_1\right]$.

Exercise 2.2 If $\mathbb{E}\left[Y \mid X\right] = a + bX$, find $\mathbb{E}\left[YX\right]$ as a function of moments of X.

Exercise 2.3 Prove Theorem 2.4.4 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 $\sigma^2(X)$ is the best predictor of e^2 given X:

(a) Write down the mean-squared error of a predictor $h(X)$ for e^2.

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

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

Exercise 2.6 Use $Y = m(X) + e$ to show that $\operatorname{var}[Y] = \operatorname{var}[m(X)] + \sigma^2$.

Exercise 2.7 Show that the conditional variance can be written as $\sigma^2(X) = \mathbb{E}\left[Y^2 \mid X\right] - (\mathbb{E}[Y \mid X])^2$.

Exercise 2.8 Suppose that Y is discrete-valued, taking values only on the non-negative integers, and the conditional distribution of Y given $X = x$ is Poisson:

$$\mathbb{P}\left[Y = j \mid X = x\right] = \frac{\exp\left(-x'\beta\right)\left(x'\beta\right)^j}{j!}, \qquad j = 0, 1, 2, ...$$

Compute $\mathbb{E}[Y \mid X]$ and $\operatorname{var}[Y \mid X]$. Does this justify a linear regression model of the form $Y = X'\beta + e$?

Hint: If $\mathbb{P}\left[Y = j\right] = \dfrac{\exp(-\lambda)\lambda^j}{j!}$ 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 $Y = X'\beta + e$ and $\mathbb{E}[e \mid X] = 0$, then e is independent of X.

Exercise 2.13 True or False. If $Y = X'\beta + e$ and $\mathbb{E}[Xe] = 0$, then $\mathbb{E}[e \mid X] = 0$.

Exercise 2.14 True or False. If $Y = X'\beta + e$, $\mathbb{E}[e \mid X] = 0$, and $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$, then e is independent of X.

Exercise 2.15 Consider the intercept-only model $Y = \alpha + e$ with α 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 expectation $m(x) = \mathbb{E}[Y \mid X = x]$. Are the best linear predictor and conditional expectation different?

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

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

Show that $\mathbb{E}\left[g\left(X, m, s\right)\right] = 0$ if and only if $m = \mu$ and $s = \sigma^2$.

Exercise 2.18 Suppose that $X = (1, X_2, X_3)$ where $X_3 = \alpha_1 + \alpha_2 X_2$ is a linear function of X_2.

(a) Show that $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is not invertible.

(b) Use a linear transformation of X to find an expression for the best linear predictor of Y given X. (Be explicit, do not just use the generalized inverse formula.)

Exercise 2.19 Show (2.47)-(2.48), namely that for

$$d(\beta) = \mathbb{E}\left[\left(m(X) - X'\beta\right)^2\right]$$

then

$$\beta = \underset{b \in \mathbb{R}^k}{\text{argmin}}\, d(b) = \left(\mathbb{E}\left[XX'\right]\right)^{-1} \mathbb{E}\left[Xm(X)\right] = \left(\mathbb{E}\left[XX'\right]\right)^{-1} \mathbb{E}\left[XY\right].$$

Hint: To show $\mathbb{E}[Xm(X)] = \mathbb{E}[XY]$ use the law of iterated expectations.

Exercise 2.20 Verify that (2.57) holds with $m(X)$ defined in (2.6) when (Y, X) have a joint density $f(y, x)$.

Exercise 2.21 Consider the short and long projections

$$Y = X\gamma_1 + e$$

$$Y = X\beta_1 + X^2\beta_2 + u$$

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

(b) Take the long projection is $Y = X\theta_1 + X^3\theta_2 + v$. Is there a condition under which $\gamma_1 = \theta_1$?

Exercise 2.22 Take the homoskedastic model

$$\begin{aligned} Y &= X_1'\beta_1 + X_2'\beta_2 + e \\ \mathbb{E}\left[e \mid X_1, X_2\right] &= 0 \\ \mathbb{E}\left[e^2 \mid X_1, X_2\right] &= \sigma^2 \\ \mathbb{E}\left[X_2 \mid X_1\right] &= \Gamma X_1. \end{aligned}$$

Assume $\Gamma \neq 0$. Suppose the parameter β_1 is of interest. We know that the exclusion of X_2 creates omited variable bias in the projection coefficient on X_2. It also changes the equation error. Our question is: what is the effect on the homoskedasticity property of the induced equation error? Does the exclusion of X_2 induce heteroskedasticity or not? Be specific.

Chapter 3

The Algebra of Least Squares

3.1 Introduction

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

3.2 Samples

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

$$\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1} \mathbb{E}\left[XY\right]. \tag{3.1}$$

We can estimate β from samples which include joint measurements of (Y, X). For example, supposing we are interested in estimating a wage equation, we would use a dataset with observations on wages (or weekly earnings), education, experience (or age), and demographic characteristics (gender, race, location). One possible dataset is the Current Population Survey (CPS), a survey of U.S. households which includes questions on employment, income, education, and demographic characteristics.

Notationally we wish to distinguish observations (realizations) from the underlying random variables. The random variables are (Y, X). The observations are (Y_i, X_i). From the vantage of the researcher the latter are numbers. From the vantage of statistical theory we view them as realizations of random variables. For individual observations we append a subscript i which runs from 1 to n, thus the i^{th} observation is (Y_i, X_i). The number n is the sample size. The **dataset** or **sample** is $\{(Y_i, X_i) : i = 1, ..., n\}$.

From the viewpoint of empirical analysis a dataset is an array of numbers. It is typically organized as a table where each column is a variable and each row is an observation. For empirical analysis the dataset is fixed in the sense that they are numbers presented to the researcher. For statistical analysis we view the dataset as random, or more precisely as a realization of a random process.

The individual observations could be draws from a common (homogeneous) distribution or could be draws from heterogeneous distributions. The simplest approach is to assume homogeneity – that the observations are realizations from an identical underlying population F.

Assumption 3.1 The variables $\{(Y_1, X_1), ..., (Y_i, X_i), ..., (Y_n, X_n)\}$ are **identically distributed**; they are draws from a common distribution F.

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

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

The linear projection model applies to the random variables (Y, X). This is the probability model described in Section 2.18. The model is

$$Y = X'\beta + e \tag{3.2}$$

where the linear projection coefficient β is defined as

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

the minimizer of the expected squared error

$$S(\beta) = \mathbb{E}\left[\left(Y - X'\beta\right)^2\right]. \tag{3.4}$$

The coefficient has the explicit solution (3.1).

3.3 Moment Estimators

We want to estimate the coefficient β defined in (3.1) from the sample of observations. Notice that β is written as a function of certain population expectations. In this context an appropriate estimator is the same function of the sample moments. Let's explain this in detail.

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

$$\mu = \mathbb{E}[Y] = \int_{-\infty}^{\infty} y dF(y). \tag{3.5}$$

The expectation μ is a function of the distribution F. To estimate μ given n random variables Y_i from F a natural estimator is the sample mean

$$\widehat{\mu} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Notice that we have written this using two pieces of notation. The notation $\overline{Y}$ with the bar on top is conventional for a sample mean. The notation $\widehat{\mu}$ with the hat "^" is conventional in econometrics to denote an estimator of the parameter μ. In this case $\overline{Y}$ is the estimator of μ, so $\widehat{\mu}$ and $\overline{Y}$ are the same. The sample mean $\overline{Y}$ can be viewed as the natural analog of the population mean (3.5) because $\overline{Y}$ equals the expectation (3.5) with respect to the empirical distribution – the discrete distribution which puts weight $1/n$ on each observation Y_i. There are other justifications for $\overline{Y}$ as an estimator for μ. We will defer these discussions for now. Suffice it to say that it is the conventional estimator.

Now suppose that we are interested in a set of population expectations of possibly nonlinear functions of a random vector Y, say $\mu = \mathbb{E}[h(Y)]$. For example, we may be interested in the first two moments of Y, $\mathbb{E}[Y]$ and $\mathbb{E}\left[Y^2\right]$. In this case the natural estimator is the vector of sample means,

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

We call $\widehat{\mu}$ the **moment estimator** for μ. For example, if $h(y) = (y, y^2)'$ then $\widehat{\mu}_1 = n^{-1}\sum_{i=1}^{n} Y_i$ and $\widehat{\mu}_2 = n^{-1}\sum_{i=1}^{n} Y_i^2$.

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

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

In general, many parameters of interest can be written as a function of moments of Y. Notationally, $\beta = g(\mu)$ and $\mu = \mathbb{E}[h(Y)]$. Here, Y are the random variables, $h(Y)$ are functions (transformations) of the random variables, and μ is the expectation of these functions. β is the parameter of interest, and is the (nonlinear) function $g(\cdot)$ of these expectations.

In this context a natural estimator of β is obtained by replacing μ with $\widehat{\mu}$. Thus $\widehat{\beta} = g\left(\widehat{\mu}\right)$. The estimator $\widehat{\beta}$ is often called a **plug-in estimator**. We also call $\widehat{\beta}$ a moment, or moment-based, estimator of β since it is a natural extension of the moment estimator $\widehat{\mu}$.

Take the example of the variance $\sigma^2 = \operatorname{var}[Y]$. Its moment estimator is

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

This is not the only possible estimator for σ^2 (there is also the well-known bias-corrected estimator) but $\widehat{\sigma}^2$ is a straightforward and simple choice.

3.4 Least Squares Estimator

The linear projection coefficient β is defined in (3.3) as the minimizer of the expected squared error $S(\beta)$ defined in (3.4). For given β, the expected squared error is the expectation of the squared error $\left(Y - X'\beta\right)^2$. The moment estimator of $S(\beta)$ is the sample average:

$$\widehat{S}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2 = \frac{1}{n}\mathrm{SSE}(\beta) \tag{3.6}$$

where

$$\mathrm{SSE}(\beta) = \sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2$$

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

Since $\widehat{S}(\beta)$ is a sample average we can interpret it as an estimator of the expected squared error $S(\beta)$. Examining $\widehat{S}(\beta)$ as a function of β is informative about how $S(\beta)$ varies with β. Since the projection coefficient minimizes $S(\beta)$ an analog estimator minimizes (3.6).

We define the estimator $\widehat{\beta}$ as the minimizer of $\widehat{S}(\beta)$.

Definition 3.1 The **least squares estimator** is $\widehat{\beta} = \underset{\beta \in \mathbb{R}^k}{\operatorname{argmin}} \widehat{S}(\beta)$ where $\widehat{S}(\beta) = \dfrac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2$.

As $\widehat{S}(\beta)$ is a scale multiple of $\mathrm{SSE}(\beta)$ we may equivalently define $\widehat{\beta}$ as the minimizer of $\mathrm{SSE}(\beta)$. Hence $\widehat{\beta}$ is commonly called the **least squares (LS)** estimator of β. The estimator is also commonly refered to as the **ordinary least squares (OLS)** estimator. For the origin of this label see the historical discussion on Adrien-Marie Legendre below. Here, as is common in econometrics, we put a hat "^" over the parameter β to indicate that $\widehat{\beta}$ is a sample estimator of β. This is a helpful convention. Just by seeing the symbol $\widehat{\beta}$ we can immediately interpret it as an estimator (because of the hat) of the parameter β. Sometimes when we want to be explicit about the estimation method, we will write $\widehat{\beta}_{\mathrm{ols}}$ to signify that it is the OLS estimator. It is also common to see the notation $\widehat{\beta}_n$, 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 β and sample estimators such as $\widehat{\beta}$. The population parameter β is a non-random feature of the population while the sample estimator $\widehat{\beta}$ is a random feature of a random sample. β is fixed, while $\widehat{\beta}$ varies across samples.

3.5 Solving for Least Squares with One Regressor

For simplicity, we start by considering the case $k = 1$ so that there is a scalar regressor X and a scalar coefficient β. To illustrate, Figure 3.1(a) displays a scatter plot[1] of 20 pairs (Y_i, X_i).

The sum of squared errors $\mathrm{SSE}(\beta)$ is a function of β. Given β we calculate the "error" $Y_i - X_i\beta$ by taking the vertical distance between Y_i and $X_i\beta$. This can be seen in Figure 3.1(a) by the vertical lines which connect the observations to the straight line. These vertical lines are the errors $Y_i - X_i\beta$. The sum of squared errors is the sum of the 20 squared lengths.

The sum of squared errors is the function

$$\mathrm{SSE}(\beta) = \sum_{i=1}^{n}\left(Y_i - X_i\beta\right)^2 = \left(\sum_{i=1}^{n} Y_i^2\right) - 2\beta\left(\sum_{i=1}^{n} X_i Y_i\right) + \beta^2\left(\sum_{i=1}^{n} X_i^2\right).$$

This is a quadratic function of β. The sum of squared error function is displayed in Figure 3.1(b) over the range $[2,4]$. The coefficient β ranges along the x-axis. The sum of squared errors $\mathrm{SSE}(\beta)$ as a function of β is displayed on the y-axis.

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 $\mathrm{SSE}(\beta)$ is

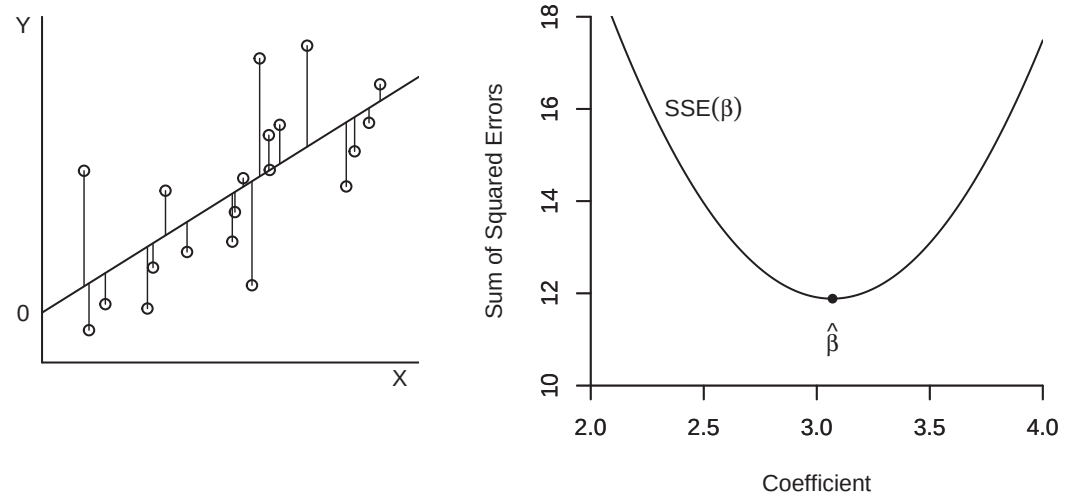
$$\widehat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}. \tag{3.7}$$

For example, the minimizer of the sum of squared error function displayed in Figure 3.1(b) is $\widehat{\beta} = 3.07$, and is marked on the x-axis.

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

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} 1\, Y_i}{\sum_{i=1}^{n} 1^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y}, \tag{3.8}$$

[1] The observations were generated by simulation as $X \sim U[0,1]$ and $Y \sim \mathrm{N}[3X, 1]$.



(a) Deviation from Fitted Line

(b) Sum of Squared Error Function

Figure 3.1: Regression With One Regressor

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 shows that the OLS estimator in the intercept-only model is the sample mean.

Technically, the estimator $\widehat{\beta}$ in (3.7) only exists if the denominator is non-zero. Since it is a sum of squares it is necessarily non-negative. Thus $\widehat{\beta}$ exists if $\sum_{i=1}^n X_i^2 > 0$.

3.6 Solving for Least Squares with Multiple Regressors

We now consider the case with $k > 1$ so that the coefficient $\beta \in \mathbb{R}^k$ is a vector.

To illustrate, Figure 3.2 displays a scatter plot of 100 triples (Y_i, X_{1i}, X_{2i}). The regression function $x'\beta = x_1\beta_1 + x_2\beta_2$ is a 2-dimensional surface and is shown as the plane in Figure 3.2.

The sum of squared errors $\text{SSE}(\beta)$ is a function of the vector β. For any β the error $Y_i - X_i'\beta$ is the vertical distance between Y_i and $X_i'\beta$. This can be seen in Figure 3.2 by the vertical lines which connect the observations to the plane. As in the single regressor case these vertical lines are the errors $e_i = Y_i - X_i'\beta$. The sum of squared errors is the sum of the 100 squared lengths.

The sum of squared errors can be written as

$$\text{SSE}(\beta) = \sum_{i=1}^n Y_i^2 - 2\beta' \sum_{i=1}^n X_i Y_i + \beta' \sum_{i=1}^n X_i X_i' \beta.$$

As in the single regressor case this is a quadratic function in β. The difference is that in the multiple regressor case this is a vector-valued quadratic function. To visualize the sum of squared errors function Figure 3.3(a) displays $\text{SSE}(\beta)$. Another way to visualize a 3-dimensional surface is by a contour plot. A contour plot of the same $\text{SSE}(\beta)$ function is shown in Figure 3.3(b). The contour lines are points in the (β_1, β_2) space where $\text{SSE}(\beta)$ takes the same value. The contour lines are elliptical because $\text{SSE}(\beta)$ is quadratic.

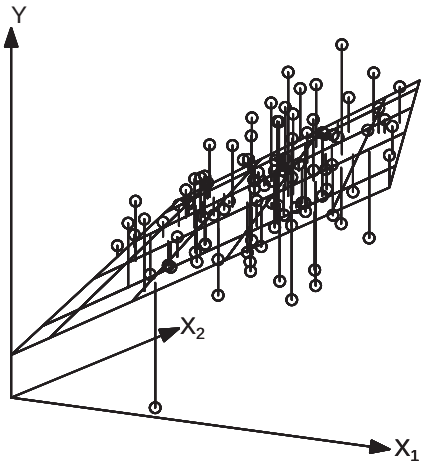


Figure 3.2: Regression with Two Variables

The least squares estimator $\widehat{\beta}$ minimizes SSE(β). A simple way to find the minimum is by solving the first-order conditions. The latter are

$$0 = \frac{\partial}{\partial \beta} \text{SSE}(\widehat{\beta}) = -2\sum_{i=1}^{n} X_i Y_i + 2\sum_{i=1}^{n} X_i X_i' \widehat{\beta}. \tag{3.9}$$

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

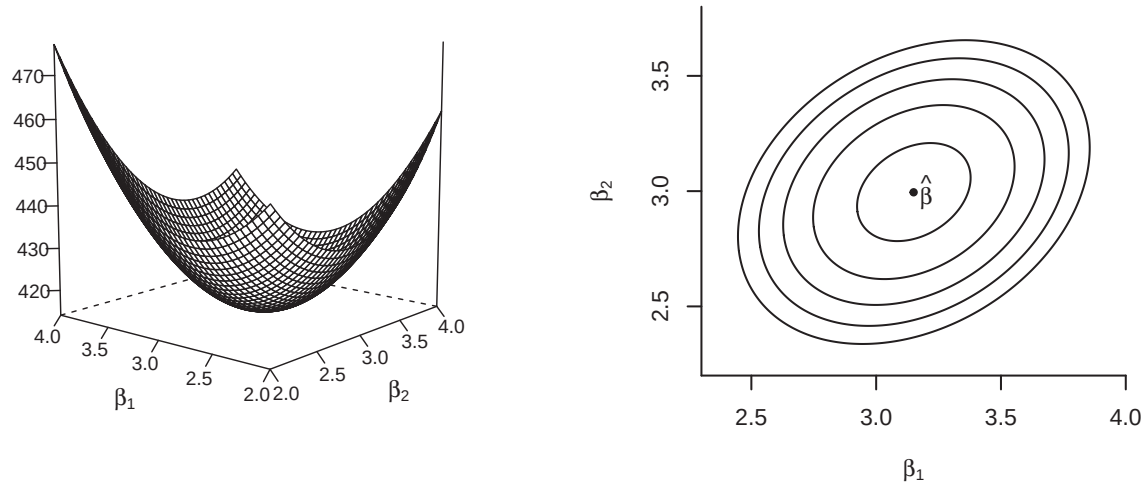
The solution for $\widehat{\beta}$ may be found by solving the system of k equations in (3.9). We can write this solution compactly using matrix algebra. Dividing (3.9) by 2 we obtain

$$\sum_{i=1}^{n} X_i X_i' \widehat{\beta} = \sum_{i=1}^{n} X_i Y_i. \tag{3.10}$$

This is a system of equations of the form $\boldsymbol{A}\boldsymbol{b} = \boldsymbol{c}$ where $\boldsymbol{A}$ is $k \times k$ and $\boldsymbol{b}$ and $\boldsymbol{c}$ are $k \times 1$. The solution is $\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{c}$, and can be obtained by pre-multiplying $\boldsymbol{A}\boldsymbol{b} = \boldsymbol{c}$ by $\boldsymbol{A}^{-1}$ and using the matrix inverse property $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_k$. Applied to (3.10) we find an explicit formula for the least squares estimator

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right). \tag{3.11}$$

This is the natural estimator of the best linear projection coefficient β defined in (3.3), and could also be called the **linear projection estimator**.



(a) Sum of Squared Error Function

(b) SSE Contour

Figure 3.3: SSE with Two Regressors

Recall that we claimed that $\widehat{\beta}$ in (3.11) is the minimizer of SSE(β), and found it by solving the first-order conditions. To be complete we should verify the second-order conditions. We calculate that

$$\frac{\partial^2}{\partial\beta\partial\beta'}\text{SSE}(\beta) = 2\sum_{i=1}^{n} X_i X_i'$$

which is a positive semi-definite matrix. If actually positive definite, then the second-order condition for minimization is satisfied, in which case $\widehat{\beta}$ is the unique minimizer of SSE(β).

Returning to the example sum of squared errors function SSE(β) displayed in Figure 3.3, the least squares estimator $\widehat{\beta}$ is the the pair $(\widehat{\beta}_1, \widehat{\beta}_2)$ which minimize this function; visually it is the low spot in the 3-dimensional graph, and is marked in Figure 3.3(b) as the center point of the contour plots.

Take equation (3.11) and suppose that $k = 1$. In this case X_i is scalar so $X_i X_i' = X_i^2$. Then (3.11) simplifies to the expression (3.7) previously derived. The expression (3.11) is a notationally simple generalization but requires a careful attention to vector and matrix manipulations.

Alternatively, equation (3.1) writes the projection coefficient β as an explicit function of the population moments $\boldsymbol{Q}_{XY}$ and $\boldsymbol{Q}_{XX}$. Their moment estimators are the sample moments

$$\widehat{\boldsymbol{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

$$\widehat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'.$$

The moment estimator of β replaces the population moments in (3.1) with the sample moments:

$$\begin{aligned}\widehat{\beta} &= \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{XY} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) \\ &= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right)\end{aligned}$$

which is identical with (3.11).

Technically, the estimator $\widehat{\beta}$ is unique and equals (3.11) only if the inverted matrix is actually invertible, which holds if (and only if) this matrix is positive definite. This excludes the case that X_i contains redundant regressors. This will be discussed further in Section 3.24.

Theorem 3.1 If $\sum_{i=1}^{n} X_i X_i' > 0$, the least squares estimator is unique and equals

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right).$$

Adrien-Marie Legendre

The method of least squares was published in 1805 by the French mathematician Adrien-Marie Legendre (1752-1833). Legendre proposed least squares as a solution to the algebraic problem of solving a system of equations when the number of equations exceeded the number of unknowns. This was a vexing and common problem in astronomical measurement. As viewed by Legendre, (3.2) is a set of n equations with k unknowns. As the equations cannot be solved exactly, Legendre's goal was to select β 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 (3.9) is a system of k equations with k unknowns which can be solved by "ordinary" methods. Hence the method became known as **Ordinary Least Squares** and to this day we still use the abbreviation OLS to refer to Legendre's estimation method.

3.7 Illustration

We illustrate the least squares estimator in practice with the data set used to calculate the estimates reported in Chapter 2. This is the March 2009 Current Population Survey, which has extensive information on the U.S. population. This data set is described in more detail in Section 3.22. For this illustration

we use the sub-sample of married (spouse present) Black female wage earners with 12 years potential work experience. This sub-sample has 20 observations[2].

In Table 3.1 we display the observations for reference. Each row is an individual observation which are the data for an individual person. The columns correspond to the variables (measurements) for the individuals. The second column is the reported *wage* (total annual earnings divided by hours worked). The third column is the natural logarithm of the wage. The fourth column is years of *education*. The fifth and six columns are further transformations, specifically the square of *education* and the product of *education* and log(*wage*). The bottom row are the sums of the elements in that column.

Table 3.1: Observations From CPS Data Set

Observation	wage	log(wage)	education	education^2	education×log(wage)
1	37.93	3.64	18	324	65.44
2	40.87	3.71	18	324	66.79
3	14.18	2.65	13	169	34.48
4	16.83	2.82	16	256	45.17
5	33.17	3.50	16	256	56.03
6	29.81	3.39	18	324	61.11
7	54.62	4.00	16	256	64.00
8	43.08	3.76	18	324	67.73
9	14.42	2.67	12	144	32.03
10	14.90	2.70	16	256	43.23
11	21.63	3.07	18	324	55.44
12	11.09	2.41	16	256	38.50
13	10.00	2.30	13	169	29.93
14	31.73	3.46	14	196	48.40
15	11.06	2.40	12	144	28.84
16	18.75	2.93	16	256	46.90
17	27.35	3.31	14	196	46.32
18	24.04	3.18	16	256	50.76
19	36.06	3.59	18	324	64.53
20	23.08	3.14	16	256	50.22
Sum	515	62.64	314	5010	995.86

Putting the variables into the standard regression notation, let Y_i be log(*wage*) and X_i be years of *education* and an intercept. Then from the column sums in Table 3.1 we have

$$\sum_{i=1}^{n} X_i Y_i = \begin{pmatrix} 995.86 \\ 62.64 \end{pmatrix}$$

and

$$\sum_{i=1}^{n} X_i X_i' = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}.$$

Taking the inverse we obtain

$$\left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} = \begin{pmatrix} 0.0125 & -0.196 \\ -0.196 & 3.124 \end{pmatrix}.$$

[2]This sample was selected specifically so that it has a small number of observations, facilitating exposition.

Thus by matrix multiplication

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

In practice the regression estimates $\widehat{\beta}$ are computed by computer software without the user taking the explicit steps listed above. However, it is useful to understand that the least squares estimator can be calculated by simple algebraic operations. If your data is in a spreadsheet similar to Table 3.1, then the listed transformations (logarithm, squares, cross-products, column sums) can be computed by spreadsheet operations. $\widehat{\beta}$ could then be calculated by matrix inversion and multiplication. Once again, this is rarely done by applied economists because computer software is available to ease the process.

We often write the estimated equation using the format

$$\widehat{\log(wage)} = 0.155\ education + 0.698. \tag{3.12}$$

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

Another use of the estimated equation (3.12) is for prediction. Suppose one individual has 12 years of education and a second has 16. Using (3.12) we find that the first's expected log wage is

$$\widehat{\log(wage)} = 0.155 \times 12 + 0.698 = 2.56$$

and for the second

$$\widehat{\log(wage)} = 0.155 \times 16 + 0.698 = 3.18.$$

Equation (3.12) is called a **bivariate regression** as there are two variables. It is also called a **simple regression** as there is a single regressor. A **multiple regression** has two or more regressors and allows a more detailed investigation. Let's take an example similar to (3.12) 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

$$\widehat{\log(wage)} = 0.143\ education + 0.036\ experience - 0.071\ experience^2/100 + 0.575. \tag{3.13}$$

These estimates suggest a 14% increase in mean wages per year of education holding experience constant.

3.8 Least Squares Residuals

As a by-product of estimation we define the **fitted value** $\widehat{Y}_i = X_i'\widehat{\beta}$ and the **residual**

$$\widehat{e}_i = Y_i - \widehat{Y}_i = Y_i - X_i'\widehat{\beta}. \tag{3.14}$$

Sometimes $\widehat{Y}_i$ is called the predicted value but this is a misleading label. The fitted value $\widehat{Y}_i$ is a function of the entire sample including Y_i, and thus cannot be interpreted as a valid prediction of Y_i. It is thus more accurate to describe $\widehat{Y}_i$ as a *fitted* rather than a *predicted* value.

Note that $Y_i = \widehat{Y}_i + \widehat{e}_i$ and

$$Y_i = X_i'\widehat{\beta} + \widehat{e}_i. \tag{3.15}$$

We make a distinction between the **error** e_i and the **residual** $\widehat{e}_i$. The error e_i is unobservable while the residual $\widehat{e}_i$ is an estimator. These two variables are frequently mislabeled which can cause confusion.

Equation (3.9) implies that

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

To see this by a direct calculation, using (3.14) and (3.11),

$$\begin{aligned}
\sum_{i=1}^{n} X_i \widehat{e}_i &= \sum_{i=1}^{n} X_i \left(Y_i - X_i' \widehat{\beta}\right) \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i X_i' \widehat{\beta} \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i X_i' \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right) \\
&= \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i Y_i = 0.
\end{aligned}$$

When X_i contains a constant an implication of (3.16) is

$$\frac{1}{n} \sum_{i=1}^{n} \widehat{e}_i = 0. \tag{3.17}$$

Thus the residuals have a sample mean of zero and the sample correlation between the regressors and the residual is zero. These are algebraic results and hold true for all linear regression estimates.

3.9 Demeaned Regressors

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

$$Y_i = X_i' \beta + \alpha + e_i$$

where α is the intercept and X_i does not contain a constant. The least squares estimates and residuals can be written as $Y_i = X_i' \widehat{\beta} + \widehat{\alpha} + \widehat{e}_i$.

In this case (3.16) can be written as the equation system

$$\begin{aligned}
\sum_{i=1}^{n} \left(Y_i - X_i' \widehat{\beta} - \widehat{\alpha}\right) &= 0 \\
\sum_{i=1}^{n} X_i \left(Y_i - X_i' \widehat{\beta} - \widehat{\alpha}\right) &= 0.
\end{aligned}$$

The first equation implies

$$\widehat{\alpha} = \overline{Y} - \overline{X}' \widehat{\beta}.$$

Subtracting from the second we obtain

$$\sum_{i=1}^{n} X_i \left(\left(Y_i - \overline{Y}\right) - \left(X_i - \overline{X}\right)' \widehat{\beta}\right) = 0.$$

Solving for $\widehat{\beta}$ we find

$$\begin{aligned}
\widehat{\beta} &= \left(\sum_{i=1}^{n} X_i \left(X_i - \overline{X}\right)'\right)^{-1} \left(\sum_{i=1}^{n} X_i \left(Y_i - \overline{Y}\right)\right) \\
&= \left(\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)'\right)^{-1} \left(\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)\right).
\end{aligned} \tag{3.18}$$

Thus the OLS estimator for the slope coefficients is OLS with demeaned data and no intercept.

The representation (3.18) is known as the demeaned formula for the least squares estimator.

3.10 Model in Matrix Notation

For many purposes, including computation, it is convenient to write the model and statistics in matrix notation. The n linear equations $Y_i = X_i'\beta + e_i$ make a system of n equations. We can stack these n equations together as

$$\begin{aligned} Y_1 &= X_1'\beta + e_1 \\ Y_2 &= X_2'\beta + e_2 \\ &\vdots \\ Y_n &= X_n'\beta + e_n. \end{aligned}$$

Define

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad \boldsymbol{X} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}, \qquad \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

Observe that $\boldsymbol{Y}$ and $\boldsymbol{e}$ are $n \times 1$ vectors and $\boldsymbol{X}$ is an $n \times k$ matrix. The system of n equations can be compactly written in the single equation

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}. \tag{3.19}$$

Sample sums can be written in matrix notation. For example

$$\begin{aligned} \sum_{i=1}^{n} X_i X_i' &= \boldsymbol{X}'\boldsymbol{X} \\ \sum_{i=1}^{n} X_i Y_i &= \boldsymbol{X}'\boldsymbol{Y}. \end{aligned}$$

Therefore the least squares estimator can be written as

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

The matrix version of (3.15) and estimated version of (3.19) is

$$\boldsymbol{Y} = \boldsymbol{X}\widehat{\beta} + \widehat{\boldsymbol{e}}.$$

Equivalently the residual vector is

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

Using the residual vector we can write (3.16) as

$$\boldsymbol{X}'\widehat{\boldsymbol{e}} = 0.$$

It can also be useful to write the sum of squared error criterion as

$$\mathrm{SSE}(\beta) = \left(\boldsymbol{Y} - \boldsymbol{X}\beta\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\beta\right).$$

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.

Theorem 3.2 Important Matrix Expressions

$$\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right)$$
$$\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}$$
$$\boldsymbol{X}'\widehat{\boldsymbol{e}} = 0.$$

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 10^{th} to 2^{nd} century BCE.

3.11 Projection Matrix

Define the matrix

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

Observe that

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

This is a property of a **projection matrix**. More generally, for any matrix $\boldsymbol{Z}$ which can be written as $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{\Gamma}$ for some matrix $\boldsymbol{\Gamma}$ (we say that $\boldsymbol{Z}$ lies in the **range space** of $\boldsymbol{X}$), then

$$\boldsymbol{P}\boldsymbol{Z} = \boldsymbol{P}\boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{X}\boldsymbol{\Gamma} = \boldsymbol{Z}.$$

As an important example, if we partition the matrix $\boldsymbol{X}$ into two matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ so that $\boldsymbol{X} = [\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$ then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$. (See Exercise 3.7.)

The projection matrix $\boldsymbol{P}$ has the algebraic property that it is **idempotent**: $\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$. See Theorem 3.3.2 below. For the general properties of projection matrices see Section A.11.

The matrix $\boldsymbol{P}$ creates the fitted values in a least squares regression:

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

Because of this property $\boldsymbol{P}$ is also known as the **hat matrix**.

A special example of a projection matrix occurs when $\boldsymbol{X} = \mathbf{1}_n$ is an n-vector of ones. Then

$$\boldsymbol{P} = \mathbf{1}_n\left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1}\mathbf{1}_n' = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'.$$

Note that in this case

$$\boldsymbol{P}\boldsymbol{Y} = \mathbf{1}_n\left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1}\mathbf{1}_n'\boldsymbol{Y} = \mathbf{1}_n\overline{Y}$$

creates an n-vector whose elements are the sample mean $\overline{Y}$.

The projection matrix $\boldsymbol{P}$ appears frequently in algebraic manipulations in least squares regression. The matrix has the following important properties.

Theorem 3.3 The projection matrix $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ for any $n \times k$ $\boldsymbol{X}$ with $n \geq k$ has the following algebraic properties.

1. $\boldsymbol{P}$ is symmetric ($\boldsymbol{P}' = \boldsymbol{P}$).
2. $\boldsymbol{P}$ is idempotent ($\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}$).
3. $\operatorname{tr}\boldsymbol{P} = k$.
4. The eigenvalues of $\boldsymbol{P}$ are 1 and 0.
5. $\boldsymbol{P}$ has k eigenvalues equalling 1 and $n-k$ equalling 0.
6. $\operatorname{rank}(\boldsymbol{P}) = k$.

We close this section by proving the claims in Theorem 3.3. Part 1 holds because

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

To establish part 2, the fact that $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$ implies that

$$\boldsymbol{P}\boldsymbol{P} = \boldsymbol{P}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' = \boldsymbol{P}$$

as claimed. For part 3,

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

See Appendix A.5 for definition and properties of the trace operator.

Appendix A.11 shows that part 4 holds for any idempotent matrix. For part 5, since $\operatorname{tr}\boldsymbol{P}$ equals the sum of the n eigenvalues and $\operatorname{tr}\boldsymbol{P} = k$ by part 3, it follows that there are k eigenvalues equalling 1 and the remainder $n-k$ equalling 0.

For part 6, observe that $\boldsymbol{P}$ is positive semi-definite because its eigenvalues are all non-negative. By Theorem A.4.5 its rank equals the number of positive eigenvalues, which is k as claimed.

3.12 Annihilator Matrix

Define

$$\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$$

where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. Note that

$$\boldsymbol{M}\boldsymbol{X} = (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{P}\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = 0. \tag{3.21}$$

Thus $\boldsymbol{M}$ and $\boldsymbol{X}$ are orthogonal. We call $\boldsymbol{M}$ the **annihilator matrix** due to the property that for any matrix $\boldsymbol{Z}$ in the range space of $\boldsymbol{X}$ then

$$\boldsymbol{M}\boldsymbol{Z} = \boldsymbol{Z} - \boldsymbol{P}\boldsymbol{Z} = 0.$$

For example, $\boldsymbol{M}\boldsymbol{X}_1 = 0$ for any subcomponent $\boldsymbol{X}_1$ of $\boldsymbol{X}$, and $\boldsymbol{M}\boldsymbol{P} = 0$ (see Exercise 3.7).

The annihilator matrix $\boldsymbol{M}$ has similar properties with $\boldsymbol{P}$, including that $\boldsymbol{M}$ is symmetric ($\boldsymbol{M}' = \boldsymbol{M}$) and idempotent ($\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$). It is thus a projection matrix. Similarly to Theorem 3.3.3 we can calculate

$$\operatorname{tr} \boldsymbol{M} = n - k. \tag{3.22}$$

(See Exercise 3.9.) One implication is that the rank of $\boldsymbol{M}$ is $n - k$.

While $\boldsymbol{P}$ creates fitted values, $\boldsymbol{M}$ creates least squares residuals:

$$\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{P}\boldsymbol{Y} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta} = \widehat{\boldsymbol{e}}. \tag{3.23}$$

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

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

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

$$\boldsymbol{M}\boldsymbol{Y} = \boldsymbol{Y} - \mathbf{1}_n \overline{Y}.$$

For simplicity we will often write the right-hand-side as $\boldsymbol{Y} - \overline{Y}$. The i^{th} element is $Y_i - \overline{Y}$, the **demeaned** value of Y_i.

We can also use (3.23) to write an alternative expression for the residual vector. Substituting $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$ into $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{Y}$ and using $\boldsymbol{M}\boldsymbol{X} = \mathbf{0}$ we find

$$\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\left(\boldsymbol{X}\beta + \boldsymbol{e}\right) = \boldsymbol{M}\boldsymbol{e} \tag{3.24}$$

which is free of dependence on the regression coefficient β.

3.13 Estimation of Error Variance

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

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \tag{3.25}$$

However, this is infeasible as e_i is not observed. In this case it is common to take a two-step approach to estimation. The residuals $\widehat{e}_i$ are calculated in the first step, and then we substitute $\widehat{e}_i$ for e_i in expression (3.25) to obtain the feasible estimator

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i^2. \tag{3.26}$$

In matrix notation, we can write (3.25) and (3.26) as $\widetilde{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e}$ and

$$\widehat{\sigma}^2 = n^{-1}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}. \tag{3.27}$$

Recall the expressions $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{e}$ from (3.23) and (3.24). Applied to (3.27) we find

$$\widehat{\sigma}^2 = n^{-1}\widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}} = n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{M}\boldsymbol{e} = n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} \tag{3.28}$$

the third equality because $\boldsymbol{M}\boldsymbol{M} = \boldsymbol{M}$.

An interesting implication is that

$$\widetilde{\sigma}^2 - \widehat{\sigma}^2 = n^{-1}\boldsymbol{e}'\boldsymbol{e} - n^{-1}\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} = n^{-1}\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e} \geq 0.$$

The final inequality holds because $\boldsymbol{P}$ is positive semi-definite and $\boldsymbol{e}'\boldsymbol{P}\boldsymbol{e}$ is a quadratic form. This shows that the feasible estimator $\widehat{\sigma}^2$ is numerically smaller than the idealized estimator (3.25).

3.14 Analysis of Variance

Another way of writing (3.23) is

$$\boldsymbol{Y} = \boldsymbol{P}\boldsymbol{Y} + \boldsymbol{M}\boldsymbol{Y} = \widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{e}}. \tag{3.29}$$

This decomposition is **orthogonal**, that is

$$\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{e}} = (\boldsymbol{P}\boldsymbol{Y})'(\boldsymbol{M}\boldsymbol{Y}) = \boldsymbol{Y}'\boldsymbol{P}\boldsymbol{M}\boldsymbol{Y} = 0. \tag{3.30}$$

It follows that

$$\boldsymbol{Y}'\boldsymbol{Y} = \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} + 2\widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{e}} + \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}} = \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{Y}} + \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$$

or

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n} \widehat{Y}_i^2 + \sum_{i=1}^{n} \widehat{e}_i^2.$$

Subtracting $\overline{Y}$ from both sides of (3.29) we obtain

$$\boldsymbol{Y} - \mathbf{1}_n\overline{Y} = \widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y} + \widehat{\boldsymbol{e}}.$$

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

$$\left(\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}\right)'\widehat{\boldsymbol{e}} = \widehat{\boldsymbol{Y}}'\widehat{\boldsymbol{e}} - \overline{Y}\mathbf{1}_n'\widehat{\boldsymbol{e}} = 0$$

under (3.17). It follows that

$$\left(\boldsymbol{Y} - \mathbf{1}_n\overline{Y}\right)'\left(\boldsymbol{Y} - \mathbf{1}_n\overline{Y}\right) = \left(\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}\right)'\left(\widehat{\boldsymbol{Y}} - \mathbf{1}_n\overline{Y}\right) + \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}$$

or

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n}\widehat{e}_i^2.$$

This is commonly called the **analysis-of-variance** formula for least squares regression.

A commonly reported statistic is the **coefficient of determination** or **R-squared**:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\widehat{e}_i^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}.$$

It is often described as "the fraction of the sample variance of Y 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.

The coefficient of determination was introduced by Wright (1921).

3.15 Projections

One way to visualize least squares fitting is as a projection operation.

Write the regressor matrix as $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2\ ...\ \boldsymbol{X}_k]$ where $\boldsymbol{X}_j$ is the j^{th} column of $\boldsymbol{X}$. The range space $\mathscr{R}(\boldsymbol{X})$ of $\boldsymbol{X}$ is the space consisting of all linear combinations of the columns $\boldsymbol{X}_1$,$\boldsymbol{X}_2$,...,$\boldsymbol{X}_k$. $\mathscr{R}(\boldsymbol{X})$ is a k dimensional surface contained in $\mathbb{R}^n$. If $k = 2$ then $\mathscr{R}(\boldsymbol{X})$ is a plane. The operator $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ projects vectors onto $\mathscr{R}(\boldsymbol{X})$. The fitted values $\widehat{\boldsymbol{Y}} = \boldsymbol{P}\boldsymbol{Y}$ are the projection of $\boldsymbol{Y}$ onto $\mathscr{R}(\boldsymbol{X})$.

To visualize examine Figure 3.4. This displays the case $n = 3$ and $k = 2$. Displayed are three vectors $\boldsymbol{Y}$, $\boldsymbol{X}_1$, and $\boldsymbol{X}_2$, which are each elements of $\mathbb{R}^3$. The plane created by $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ is the range space $\mathscr{R}(\boldsymbol{X})$. Regression fitted values are linear combinations of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and so lie on this plane. The fitted value $\widehat{\boldsymbol{Y}}$ is the vector on this plane closest to $\boldsymbol{Y}$. The residual $\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ is the difference between the two. The angle between the vectors $\widehat{\boldsymbol{Y}}$ and $\widehat{\boldsymbol{e}}$ is 90°, and therefore they are orthogonal as shown.

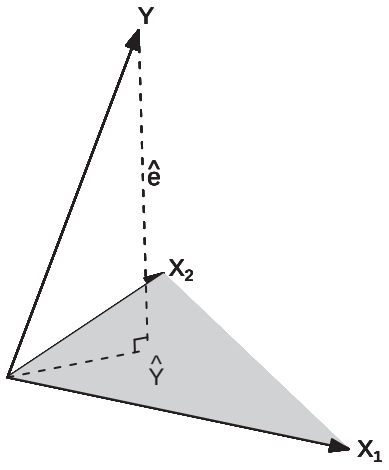


Figure 3.4: Projection of $\boldsymbol{Y}$ onto $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$

3.16 Regression Components

Partition $\boldsymbol{X} = [\boldsymbol{X}_1 \quad \boldsymbol{X}_2]$ and $\beta = (\beta_1, \beta_2)$. The regression model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{X}_2\beta_2 + \boldsymbol{e}. \tag{3.31}$$

The OLS estimator of $\beta = (\beta_1', \beta_2')'$ is obtained by regression of $\boldsymbol{Y}$ on $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ and can be written as

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

We are interested in algebraic expressions for $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

Let's first focus on $\widehat{\beta}_1$. The least squares estimator by definition is found by the joint minimization

$$\left(\widehat{\beta}_1, \widehat{\beta}_2\right) = \underset{\beta_1, \beta_2}{\operatorname{argmin}} \operatorname{SSE}\left(\beta_1, \beta_2\right) \tag{3.33}$$

where

$$\operatorname{SSE}\left(\beta_1, \beta_2\right) = \left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1 - \boldsymbol{X}_2\beta_2\right)'\left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1 - \boldsymbol{X}_2\beta_2\right).$$

An equivalent expression for $\widehat{\beta}_1$ can be obtained by concentration (nested minimization). The solution (3.33) can be written as

$$\widehat{\beta}_1 = \underset{\beta_1}{\operatorname{argmin}} \left(\min_{\beta_2} \operatorname{SSE}\left(\beta_1, \beta_2\right)\right). \tag{3.34}$$

The inner expression $\min_{\beta_2} \operatorname{SSE}\left(\beta_1, \beta_2\right)$ minimizes over β_2 while holding β_1 fixed. It is the lowest possible sum of squared errors given β_1. The outer minimization $\operatorname{argmin}_{\beta_1}$ finds the coefficient β_1 which minimizes the "lowest possible sum of squared errors given β_1". This means that $\widehat{\beta}_1$ as defined in (3.33) and (3.34) are algebraically identical.

Examine the inner minimization problem in (3.34). This is simply the least squares regression of $\boldsymbol{Y} - \boldsymbol{X}_1\beta_1$ on $\boldsymbol{X}_2$. This has solution

$$\underset{\beta_2}{\operatorname{argmin}} \operatorname{SSE}\left(\beta_1, \beta_2\right) = \left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1\right)\right)$$

with residuals

$$\begin{aligned}\boldsymbol{Y} - \boldsymbol{X}_1\beta_1 - \boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1\right)\right) &= \left(\boldsymbol{M}_2\boldsymbol{Y} - \boldsymbol{M}_2\boldsymbol{X}_1\beta_1\right) \\ &= \boldsymbol{M}_2\left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1\right)\end{aligned}$$

where

$$\boldsymbol{M}_2 = \boldsymbol{I}_n - \boldsymbol{X}_2\left(\boldsymbol{X}_2'\boldsymbol{X}_2\right)^{-1}\boldsymbol{X}_2' \tag{3.35}$$

is the annihilator matrix for $\boldsymbol{X}_2$. This means that the inner minimization problem (3.34) has minimized value

$$\begin{aligned}\min_{\beta_2} \operatorname{SSE}\left(\beta_1, \beta_2\right) &= \left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1\right)'\boldsymbol{M}_2\boldsymbol{M}_2\left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1\right) \\ &= \left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1\right)'\boldsymbol{M}_2\left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1\right)\end{aligned}$$

where the second equality holds because $\boldsymbol{M}_2$ is idempotent. Substituting this into (3.34) we find

$$\begin{aligned}\widehat{\beta}_1 &= \underset{\beta_1}{\operatorname{argmin}}\left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1\right)'\boldsymbol{M}_2\left(\boldsymbol{Y} - \boldsymbol{X}_1\beta_1\right) \\ &= \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{Y}\right).\end{aligned}$$

By a similar argument we find

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

where

$$\boldsymbol{M}_1 = \boldsymbol{I}_n - \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1' \tag{3.36}$$

is the annihilator matrix for $\boldsymbol{X}_1$.

Theorem 3.4 The least squares estimator $\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ for (3.32) has the algebraic solution

$$\widehat{\beta}_1 = \left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1'\boldsymbol{M}_2\boldsymbol{Y}\right) \tag{3.37}$$

$$\widehat{\beta}_2 = \left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{Y}\right) \tag{3.38}$$

where $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are defined in (3.36) and (3.35), respectively.

3.17 Regression Components (Alternative Derivation)*

An alternative proof of Theorem 3.4 uses an algebraic argument based on the population calculations from Section 2.22. Since this is a classic derivation we present it here for completeness.

Partition $\widehat{\boldsymbol{Q}}_{XX}$ 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 $\widehat{\boldsymbol{Q}}_{XY}$ as

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

By the partitioned matrix inversion formula (A.3)

$$\widehat{\boldsymbol{Q}}_{XX}^{-1} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11} & \widehat{\boldsymbol{Q}}_{12} \\ \widehat{\boldsymbol{Q}}_{21} & \widehat{\boldsymbol{Q}}_{22} \end{bmatrix}^{-1} \overset{\text{def}}{=} \begin{bmatrix} \widehat{\boldsymbol{Q}}^{11} & \widehat{\boldsymbol{Q}}^{12} \\ \widehat{\boldsymbol{Q}}^{21} & \widehat{\boldsymbol{Q}}^{22} \end{bmatrix} = \begin{bmatrix} \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1} & -\widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1} \\ -\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1} & \widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1} \end{bmatrix} \tag{3.39}$$

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

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

Now

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

and

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

Equation (3.38) follows.

Similarly to the calculation for $\widehat{\boldsymbol{Q}}_{11\cdot 2}$ and $\widehat{\boldsymbol{Q}}_{1Y\cdot 2}$ you can show that $\widehat{\boldsymbol{Q}}_{2Y\cdot 1} = \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{Y}$ and $\widehat{\boldsymbol{Q}}_{22\cdot 1} = \frac{1}{n}\boldsymbol{X}_2'\boldsymbol{M}_1\boldsymbol{X}_2$. This establishes (3.37). Together, this is Theorem 3.4.

3.18 Residual Regression

As first recognized by Frisch and Waugh (1933) and extended by Lovell (1963), expressions (3.37) and (3.38) can be used to show that the least squares estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ can be found by a two-step regression procedure.

Take (3.38). Since $\boldsymbol{M}_1$ is idempotent, $\boldsymbol{M}_1 = \boldsymbol{M}_1\boldsymbol{M}_1$ and thus

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

where $\widetilde{\boldsymbol{X}}_2 = \boldsymbol{M}_1\boldsymbol{X}_2$ and $\widetilde{\boldsymbol{e}}_1 = \boldsymbol{M}_1\boldsymbol{Y}$.

Thus the coefficient estimator $\widehat{\beta}_2$ is algebraically equal to the least squares regression of $\widetilde{\boldsymbol{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$. Notice that these two are $\boldsymbol{Y}$ and $\boldsymbol{X}_2$, respectively, premultiplied by $\boldsymbol{M}_1$. But we know that pre-multiplication by $\boldsymbol{M}_1$ creates least squares residuals. Therefore $\widetilde{\boldsymbol{e}}_1$ is simply the least squares residual from a regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1$, and the columns of $\widetilde{\boldsymbol{X}}_2$ are the least squares residuals from the regressions of the columns of $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$.

We have proven the following theorem.

Theorem 3.5 Frisch-Waugh-Lovell (FWL)
In the model (3.31), the OLS estimator of β_2 and the OLS residuals $\widehat{\boldsymbol{e}}$ may be computed by either the OLS regression (3.32) or via the following algorithm:

1. Regress $\boldsymbol{Y}$ on $\boldsymbol{X}_1$, obtain residuals $\widetilde{\boldsymbol{e}}_1$;
2. Regress $\boldsymbol{X}_2$ on $\boldsymbol{X}_1$, obtain residuals $\widetilde{\boldsymbol{X}}_2$;
3. Regress $\widetilde{\boldsymbol{e}}_1$ on $\widetilde{\boldsymbol{X}}_2$, obtain OLS estimates $\widehat{\beta}_2$ and residuals $\widehat{\boldsymbol{e}}$.

In some contexts (such as panel data models, to be introduced in Chapter 17), the FWL theorem can be used to greatly speed computation.

The FWL theorem is a direct analog of the coefficient representation obtained in Section 2.23. The result obtained in that section concerned the population projection coefficients; the result obtained here

concern the least squares estimators. The key message is the same. In the least squares regression (3.32) the estimated coefficient $\widehat{\beta}_2$ algebraically equals the regression of $\boldsymbol{Y}$ on the regressors $\boldsymbol{X}_2$ after the regressors $\boldsymbol{X}_1$ have been linearly projected out. Similarly, the coefficient estimate $\widehat{\beta}_1$ algebraically 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 insightful when interpreting regression coefficients.

A common application of the FWL theorem is the demeaning formula for regression obtained in (3.18). Partition $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ where $\boldsymbol{X}_1 = \mathbf{1}_n$ is a vector of ones and $\boldsymbol{X}_2$ is a matrix of observed regressors. In this case $\boldsymbol{M}_1 = \boldsymbol{I}_n - \mathbf{1}_n \left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1} \mathbf{1}_n'$. 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 $\widehat{\beta}_2$ is the OLS estimate from a regression of $Y_i - \overline{Y}$ on $X_{2i} - \overline{X}_2$:

$$\widehat{\beta}_2 = \left(\sum_{i=1}^{n} \left(X_{2i} - \overline{X}_2\right)\left(X_{2i} - \overline{X}_2\right)'\right)^{-1} \left(\sum_{i=1}^{n} \left(X_{2i} - \overline{X}_2\right)\left(Y_i - \overline{Y}\right)\right).$$

This is (3.18).

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.

3.19 Leverage Values

The **leverage** values for the regressor matrix $\boldsymbol{X}$ are the diagonal elements of the projection matrix $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. There are n leverage values, and are typically written as h_{ii} for $i = 1, ..., n$. Since

$$\boldsymbol{P} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}$$

they are

$$h_{ii} = X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i. \tag{3.40}$$

The leverage value h_{ii} is a normalized length of the observed regressor vector X_i. They appear frequently in the algebraic and statistical analysis of least squares regression, including leave-one-out regression, influential observations, robust covariance matrix estimation, and cross-validation.

A few properties of the leverage values are now listed.

Theorem 3.6

1. $0 \leq h_{ii} \leq 1$.
2. $h_{ii} \geq 1/n$ if X includes an intercept.
3. $\sum_{i=1}^{n} h_{ii} = k$.

We prove Theorem 3.6 below.

The leverage value h_{ii} measures how unusual the i^{th} observation X_i is relative to the other observations in the sample. A large h_{ii} occurs when X_i is quite different from the other sample values. A measure of overall unusualness is the maximum leverage value

$$\overline{h} = \max_{1 \leq i \leq n} h_{ii}. \tag{3.41}$$

It is common to say that a regression design is **balanced** when the leverage values are all roughly equal to one another. From Theorem 3.6.3 we deduce that complete balance occurs when $h_{ii} = \overline{h} = k/n$. An example of complete balance is when the regressors are all orthogonal dummy variables, each of which have equal occurrance of 0's and 1's.

A regression design is **unbalanced** if some leverage values are highly unequal from the others. The most extreme case is $\overline{h} = 1$. An example where this occurs is when there is a dummy regressor which takes the value 1 for only one observation in the sample.

The maximal leverage value (3.41) will change depending on the choice of regressors. For example, consider equation (3.13), the wage regression for single Asian men which has $n = 268$ observations. This regression has $\overline{h} = 0.33$. If the squared experience regressor is omitted the leverage drops to $\overline{h} = 0.10$. If a cubic in experience is added it increases to $\overline{h} = 0.76$. And if a fourth and fifth power are added it increases to $\overline{h} = 0.99$.

Some inference procedures (such as robust covariance matrix estimation and cross-validation) are sensitive to high leverage values. We will return to these issues later.

We now prove Theorem 3.6. For part 1 let s_i be an $n \times 1$ unit vector with a 1 in the i^{th} place and zeros elsewhere so that $h_{ii} = s_i' \boldsymbol{P} s_i$. Then applying the Quadratic Inequality (B.18) and Theorem 3.3.4,

$$h_{ii} = s_i' \boldsymbol{P} s_i \leq s_i' s_i \lambda_{\max}(\boldsymbol{P}) = 1$$

as claimed.

For part 2 partition $X_i = (1, Z_i')'$. Without loss of generality we can replace Z_i with the demeaned values $Z_i^* = Z_i - \overline{Z}$. Then since Z_i^* and the intercept are orthgonal

$$\begin{aligned} h_{ii} &= (1, Z_i^{*\prime}) \begin{bmatrix} n & 0 \\ 0 & \boldsymbol{Z}^{*\prime}\boldsymbol{Z}^* \end{bmatrix}^{-1} \begin{pmatrix} 1 \\ Z_i^* \end{pmatrix} \\ &= \frac{1}{n} + Z_i^{*\prime} \left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^*\right)^{-1} Z_i^* \geq \frac{1}{n}. \end{aligned}$$

For part 3, $\sum_{i=1}^{n} h_{ii} = \operatorname{tr} \boldsymbol{P} = k$ where the second equality is Theorem 3.3.3.

3.20 Leave-One-Out Regression

There are a number of statistical procedures – residual analysis, jackknife variance estimation, cross-validation, two-step estimation, hold-out sample evaluation – which make use of estimators constructed on sub-samples. Of particular importance is the case where we exclude a single observation and then repeat this for all observations. This is called **leave-one-out** (LOO) regression.

Specifically, the leave-one-out estimator of the regression coefficient β is the least squares estimator constructed using the full sample excluding a single observation i. This can be written as

$$\begin{aligned}\widehat{\beta}_{(-i)} &= \left(\sum_{j\neq i} X_j X_j'\right)^{-1}\left(\sum_{j\neq i} X_j Y_j\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X} - X_i X_i'\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y} - X_i Y_i\right) \\ &= \left(\boldsymbol{X}_{(-i)}'\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}_{(-i)}'\boldsymbol{Y}_{(-i)}. \end{aligned} \tag{3.42}$$

Here, $\boldsymbol{X}_{(-i)}$ and $\boldsymbol{Y}_{(-i)}$ are the data matrices omitting the i^{th} row. The notation $\widehat{\beta}_{(-i)}$ or $\widehat{\beta}_{-i}$ is commonly used to denote an estimator with the i^{th} observation omitted. There is a leave-one-out estimator for each observation, $i = 1, ..., n$, so we have n such estimators.

The leave-one-out predicted value for Y_i is $\widetilde{Y}_i = X_i'\widehat{\beta}_{(-i)}$. This is the predicted value obtained by estimating β on the sample without observation i and then using the covariate vector X_i to predict Y_i. Notice that $\widetilde{Y}_i$ is an authentic prediction as Y_i is not used to construct $\widetilde{Y}_i$. This is in contrast to the fitted values $\widehat{Y}_i$ which are functions of Y_i.

The **leave-one-out residual**, **prediction error**, or **prediction residual** is $\widetilde{e}_i = Y_i - \widetilde{Y}_i$. The prediction errors may be used as estimators of the errors instead of the residuals. The prediction errors are better estimators than the residuals because the former are based on authentic predictions.

The leave-one-out formula (3.42) gives the unfortunate impression that the leave-one-out coefficients and errors are computationally cumbersome, requiring n separate regressions. In the context of linear regression this is fortunately not the case. There are simple linear expressions for $\widehat{\beta}_{(-i)}$ and $\widetilde{e}_i$.

Theorem 3.7 The leave-one-out estimator and prediction error equal

$$\widehat{\beta}_{(-i)} = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \widetilde{e}_i \tag{3.43}$$

and

$$\widetilde{e}_i = (1 - h_{ii})^{-1}\,\widehat{e}_i \tag{3.44}$$

where h_{ii} are the leverage values as defined in (3.40).

We prove Theorem 3.7 at the end of the section.

Equation (3.43) shows that the leave-one-out coefficients can be calculated by a simple linear operation and do not need to be calculated using n separate regressions. Another interesting feature of equation (3.44) is that the prediction errors $\widetilde{e}_i$ are a simple scaling of the least squares residuals $\widehat{e}_i$ with the scaling dependent on the leverage values h_{ii}. If h_{ii} is small then $\widetilde{e}_i \simeq \widehat{e}_i$. However if h_{ii} is large then $\widetilde{e}_i$ can be quite different from $\widehat{e}_i$. Thus the difference between the residuals and predicted values depends on the leverage values, that is, how unusual is X_i.

To write (3.44) in vector notation, define

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

Then (3.44) is equivalent to

$$\widetilde{\boldsymbol{e}} = \boldsymbol{M}^* \widehat{\boldsymbol{e}}. \tag{3.45}$$

One use of the prediction errors is to estimate the out-of-sample mean squared error:

$$\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widetilde{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n} (1-h_{ii})^{-2} \widehat{e}_i^2. \tag{3.46}$$

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

We complete the section with a proof of Theorem 3.7. The leave-one-out estimator (3.42) can be written as

$$\widehat{\beta}_{(-i)} = \left(\boldsymbol{X}'\boldsymbol{X} - X_i X_i'\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y} - X_i Y_i\right). \tag{3.47}$$

Multiply (3.47) by $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{X} - X_i X_i'\right)$. We obtain

$$\widehat{\beta}_{(-i)} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i X_i' \widehat{\beta}_{(-i)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y} - X_i Y_i\right) = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i Y_i.$$

Rewriting

$$\widehat{\beta}_{(-i)} = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i\left(Y_i - X_i'\widehat{\beta}_{(-i)}\right) = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \widetilde{e}_i$$

which is (3.43). Premultiplying this expression by X_i' and using definition (3.40) we obtain

$$X_i'\widehat{\beta}_{(-i)} = X_i'\widehat{\beta} - X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \widetilde{e}_i = X_i'\widehat{\beta} - h_{ii}\widetilde{e}_i.$$

Using the definitions for $\widehat{e}_i$ and $\widetilde{e}_i$ we obtain $\widetilde{e}_i = \widehat{e}_i + h_{ii}\widetilde{e}_i$. Re-writing we obtain (3.44).

3.21 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 3.5 which shows a scatter plot of realizations (Y_i, X_i). The 25 observations shown with the open circles are generated by $X_i \sim U[1,10]$ and $Y_i \sim \text{N}(X_i, 4)$. 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 fitted line towards the 26^{th} observation. In fact, the slope coefficient decreases from 0.97 (which is close to the true value of 1.00) to 0.56, which is substantially reduced. Neither Y_{26} nor X_{26} are unusual values relative to their marginal distributions so this outlier would not have been detected from examination of the marginal distributions of the data. The change in the slope coefficient of -0.41 is meaningful and should raise concern to an applied economist.

From (3.43) we know that

$$\widehat{\beta} - \widehat{\beta}_{(-i)} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \widetilde{e}_i. \tag{3.48}$$

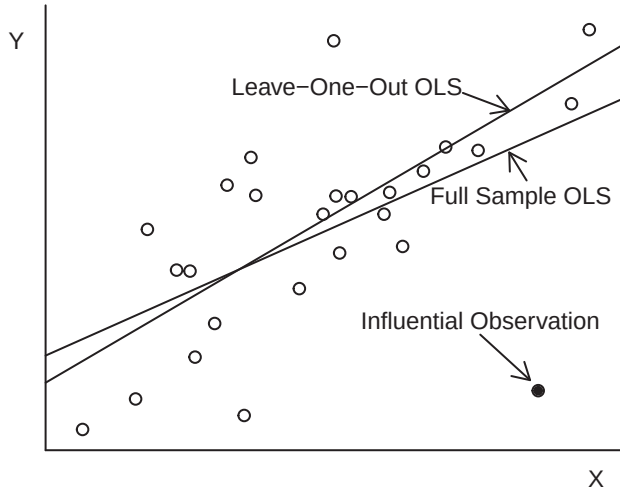


Figure 3.5: Impact of an Influential Observation on the Least-Squares Estimator

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

$$\widehat{Y}_i - \widetilde{Y}_i = X_i'\widehat{\beta} - X_i'\widehat{\beta}_{(-i)} = X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i\widetilde{e}_i = h_{ii}\widetilde{e}_i$$

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

One way to think about this is that a large leverage value h_{ii} gives the potential for observation i to be influential. A large h_{ii} means that observation i is unusual in the sense that the regressor X_i is far from its sample mean. We call an observation with large h_{ii} a **leverage point**. A leverage point is not necessarily influential as the latter also requires that the prediction error $\widetilde{e}_i$ is large.

To determine if any individual observations are influential in this sense several diagnostics have been proposed (some names include DFITS, Cook's Distance, and Welsch Distance). Unfortunately, from a statistical perspective it is difficult to recommend these diagnostics for applications as they are not based on statistical theory. Probably the most relevant measure is the change in the coefficient estimates given in (3.48). The ratio of these changes to the coefficient's standard error is called its DFBETA, and is a postestimation diagnostic available in Stata. While there is no magic threshold, the concern is whether or not an individual observation meaningfully changes an estimated coefficient of interest. A simple

diagnostic for influential observations is to calculate

$$\text{Influence} = \max_{1\leq i\leq n}\left|\widehat{Y}_i - \widetilde{Y}_i\right| = \max_{1\leq i\leq n}|h_{ii}\widetilde{e}_i|.$$

This is the largest (absolute) change in the predicted value due to a single observation. If this diagnostic is large relative to the distribution of Y 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 error, the influential observations should be examined for evidence that the observation was mis-recorded. Perhaps the observation falls outside of permitted ranges, or some observables are inconsistent (for example, a person is listed as having a job but receives earnings of $0). If it is determined that an observation is incorrectly recorded, then the observation is typically deleted from the sample. This process is often called "cleaning the data". The decisions made in this process involve a fair amount of individual judgment. [When this is done the proper practice is to retain the source data in its original form and create a program file which executes all cleaning operations (for example deletion of individual observations). The cleaned data file can be saved at this point, and then used for the subsequent statistical analysis. The point of retaining the source data and a specific program file which cleans the data is twofold: so that all decisions are documented, and so that modifications can be made in revisions and future research.] 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. This latter 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 (3.13) for single Asian men. 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 average 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. It is a moderately large leverage value, meaning that the regressor X_i is somewhat unusual. Examining further, we find that this individual is 65 years old with 8 years education, so that his potential work experience is 51 years. This is the highest experience in the subsample – the next highest is 41 years. The large leverage is due to his unusual characteristics (very low education and very high experience) within this sample. Essentially, regression (3.13) is attempting to estimate the conditional mean at *experience*= 51 with only one observation. It is not surprising that this observation determines the fit and is thus influential. A reasonable conclusion is the regression function can only be estimated over a smaller range of *experience*. We restrict the sample to individuals with less than 45 years experience, re-estimate, and obtain the following estimates.

$$\widehat{\log(wage)} = 0.144\ education + 0.043\ experience - 0.095\ experience^2/100 + 0.531. \tag{3.49}$$

For this regression, we calculate that Influence = 0.11, which is greatly reduced relative to the regression (3.13). Comparing (3.49) with (3.13), the slope coefficient for education is essentially unchanged, but the coefficients on experience and its square have slightly increased.

By eliminating the influential observation equation (3.49) can be viewed as a more robust estimate of the conditional mean for most levels of *experience*. Whether to report (3.13) or (3.49) in an application is largely a matter of judgment.

3.22 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/program-surveys/cps.html`.

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 set `cps09mar`. This data set, and all others used in this textbook, are available at `http://www.ssc.wisc.edu/~bhansen/econometrics/`.

3.23 Numerical Computation

Modern econometric estimation involves large samples and many covariates. Consequently, calculation of even simple statistics such as the least squares estimator requires a large number (millions) of arithmetic operations. In practice most economists don't need to think much about this as it is done swiftly and effortlessly on personal computers. Nevertheless it is useful to understand the underlying calculation methods as choices can occasionally make substantive differences.

While today nearly all statistical computations are made using statistical software running on electronic computers, this was not always the case. In the nineteenth and early twentieth centures "computer" was a job label for workers who made computations by hand. Computers were employed by astronomers and statistical laboratories. This fascinating job (and the fact that most computers employed in laboratories were women) has entered popular culture. For example the lives of several computers who worked for the early U.S. space program is described in the book and popular movie *Hidden Figures*, a fictional computer/astronaut is the protagonist of the novel *The Calculating Stars*, and the life of computer/astronomer Henrietta Swan Leavitt is dramatized in the play *Silent Sky*.

Until programmable electronic computers became available in the 1960s economics graduate students were routinely employed as computers. Sample sizes were considerably smaller than those seen today, but still the effort required to calculate by hand a regression with even $n = 100$ observations and $k = 5$ variables is considerable! If you are a current graduate student you should feel fortunate that the profession has moved on from the era of human computers! (Now research assistants do more elevated tasks such as writing Stata, R, and MATLAB code.)

To obtain the least squares estimate $\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right)$ we need to either invert $\boldsymbol{X}'\boldsymbol{X}$ or solve a system of equations. To be specific, let $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{c} = \boldsymbol{X}'\boldsymbol{Y}$ so that the least squares estimate can be written as either the solution to

$$\boldsymbol{A}\widehat{\beta} = \boldsymbol{c} \tag{3.50}$$

or as

$$\widehat{\beta} = \boldsymbol{A}^{-1}\boldsymbol{c}. \tag{3.51}$$

The equations (3.50) and (3.51) are algebraically identical but they suggest two distinct numerical approaches to obtain $\widehat{\beta}$. (3.50) suggests solving a system of k equations. (3.51) suggests finding $\boldsymbol{A}^{-1}$ and then multiplying by $\boldsymbol{c}$. While the two expressions are algebraically identical the implied numerical approaches are different.

In a nutshell, solving the system of equations (3.50) is numerically preferred to the matrix inversion problem (3.51). Directly solving (3.50) is faster and produces a solution with a higher degree of numerical accuracy. Thus (3.50) is generally recommended over (3.51). However, in most practical applications the choice will not make any practical difference. Contexts where the choice may make a difference is when the matrix $\boldsymbol{A}$ is ill-conditioned (to be discussed in Section 3.24) or of extremely high dimension.

Numerical methods to solve the system of equations (3.50) and calculate $\boldsymbol{A}^{-1}$ are discussed in Sections A.18 and A.19, respectively.

Statistical packages use a variety of matrix methods to solve (3.50). Stata uses the sweep algorithm which is a variant of the Gauss-Jordan algorithm discussed in Section A.18. (For the sweep algorithm see Goodnight (1979).) In R, `solve(A,b)` uses the QR decomposition. In MATLAB, `A\b` uses the Cholesky decomposition when $\boldsymbol{A}$ is positive definite and the QR decomposition otherwise.

3.24 Collinearity Errors

For the least squares estimator to be uniquely defined the regressors cannot be linearly dependent. However, it is quite easy to *attempt* to calculate a regression with linearly dependent regressors. This can occur for many reasons, including the following.

1. Including the same regressor twice.
2. Including regressors which are a linear combination of one another, such as *education*, *experience* and *age* in the CPS data set example (recall, *experience* is defined as *age-education-6*).
3. Including a dummy variable and its square.
4. Estimating a regression on a sub-sample for which a dummy variable is either all zeros or all ones.
5. Including a dummy variable interaction which yields all zeros.
6. Including more regressors than observations.

In any of the above cases the regressors are linearly dependent so $\boldsymbol{X}'\boldsymbol{X}$ is singular and the least squares estimator is not unique. If you attempt to estimate the regression, you are likely to encounter an error message. (A possible exception is MATLAB using "`A\b`", as discussed below.) The message may be that "system is exactly singular", "system is computationally singular", a variable is "omitted because of collinearity", or a coefficient is listed as "NA". In some cases (such as estimation in R using explicit matrix computation or MATLAB using the `regress` command) the program will stop execution. In other cases the program will continue to run. In Stata (and in the `lm` package in R), a regression will be reported but one or more variables will be omitted.

If any of these warnings or error messages appear, the correct response is to stop and examine the regression coding and data. Did you make an unintended mistake? Have you included a linearly dependent regressor? Are you estimating on a subsample for which the variables (in particular dummy variables) have no variation? If you can determine that one of these scenarios caused the error, the solution is immediately apparent. You need to respecify your model (either sample or regressors) so that the redundancy is eliminated. All empirical researchers encounter this error in the course of empirical work. You should not, however, simply accept output if the package has selected variables for omission. It is the researcher's job to understand the underlying cause and enact a suitable remedy.

There is also a possibility that the statistical package will not detect and report the matrix singularity. If you compute in MATLAB using explicit matrix operations and use the recommended `A\b` command to

compute the least squares estimator MATLAB may return a numerical solution without an error message even when the regressors are algebraically dependent. It is therefore recommended that you perform a numerical check for matrix singularity when using explicit matrix operations in MATLAB.

How can we numerically check if a matrix $\boldsymbol{A}$ is singular? A standard diagnostic is the **reciprocal condition number**

$$C = \frac{\lambda_{\min}(\boldsymbol{A})}{\lambda_{\max}(\boldsymbol{A})}.$$

If $C = 0$ then $\boldsymbol{A}$ is singular. If $C = 1$ then $\boldsymbol{A}$ is perfectly balanced. If C is extremely small we say that $\boldsymbol{A}$ is **ill-conditioned**. The reciprocal condition number can be calculated in MATLAB or R by the `rcond` command. Unfortunately, there is no accepted tolerance for how small C should be before regarding $\boldsymbol{A}$ as numerically singular, in part since `rcond(A)` can return a positive (but small) result even if $\boldsymbol{A}$ is algebraically singular. However, in double precision (which is typically used for computation) numerical accuracy is bounded by $2^{-52} \simeq$ 2e-16, suggesting the minimum bound $C \geq$ 2e-16.

Checking for numerical singularity is complicated by the fact that low values of C can also be caused by unbalanced or highly correlated regressors.

To illustrate, consider a wage regression using the sample from (3.13) on powers of experience X from 1 through k (e.g. $X, X^2, X^3, ..., X^k$). We calculated the reciprocal condition number C for each k, and found that C is decreasing as k increases, indicating increasing ill-conditioning. Indeed, for $k =$ 5, we find $C =$ 6e-17, which is lower than double precision accuracy. This means that a regression on (X, X^2, X^3, X^4, X^5) is ill-conditioned. The regressor matrix, however, is not singular. The low value of C is not due to algebraic singularity but rather is due to a lack of balance and high collinearity.

Ill-conditioned regressors have the potential problem that the numerical results (the reported coefficient estimates) will be inaccurate. It may not be a concern in most applications as this only occurs in extreme cases. Nevertheless, we should try and avoid ill-conditioned regressions whenever possible.

There are strategies which can reduce or even eliminate ill-conditioning. Often it is sufficient to rescale the regressors. A simple rescaling which often works for non-negative regressors is to divide each by its sample mean, thus replace X_{ji} with $X_{ji}/\overline{X}_j$. In the above example with the powers of experience, this means replacing X_i^2 with $X_i^2/\left(n^{-1}\sum_{i=1}^n X_i^2\right)$, etc. Doing so dramatically reduces the ill-conditioning. With this scaling, regressions for $k \leq 11$ satisfy $C \geq$ 1e-15. Another rescaling specific to a regression with powers is to first rescale the regressor to lie in $[-1, 1]$ before taking powers. With this scaling, regressions for $k \leq 16$ satisfy $C \geq$ 1e-15. A simpler scaling option is to rescale the regressor to lie in $[0, 1]$ before taking powers. With this scaling, regressions for $k \leq 9$ satisfy $C \geq$ 1e-15. This is often sufficient for applications.

Ill-conditioning can often be completely eliminated by orthogonalization of the regressors. This is achieved by sequentially regressing each variable (each column in $\boldsymbol{X}$) on the preceeding variables (each preceeding column), taking the residual, and then rescaling to have a unit variance. This will produce regressors which algebraically satisfy $\boldsymbol{X}'\boldsymbol{X} = n\boldsymbol{I}_n$ and have a condition number of $C = 1$. If we apply this method to the above example, we obtain a condition number close to 1 for $k \leq 20$.

What this shows is that when a regression has a small condition number it is important to examine the specification carefully. It is possible that the regressors are linearly dependent in which case one or more regressors will need to be omitted. It is also possible that the regressors are badly scaled in which case it may be useful to rescale some of the regressors. It is also possible that the variables are highly collinear in which case a possible solution is orthogonalization. These choices should be made by the researcher not by an automated software program.

3.25 Programming

Most packages allow both interactive programming (where you enter commands one-by-one) and batch programming (where you run a pre-written sequence of commands from a file). Interactive programming can be useful for exploratory analysis but eventually all work should be executed in batch mode. This is the best way to control and document your work.

Batch programs are text files where each line executes a single command. For Stata, this file needs to have the filename extension ".do", and for MATLAB ".m". For R there is no specific naming requirements, though it is typical to use the extension ".r". When writing batch files it is useful to include comments for documentation and readability. To execute a program file you type a command within the program.

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

MATLAB: `run chapter3` executes the file `chapter3.m`.

R: `source('chapter3.r')` executes the file `chapter3.r`.

There are similarities and differences between the commands used in these packages. For example:

1. Different symbols are used to create comments. * in Stata, # in R, and % in MATLAB.
2. MATLAB uses the symbol ; to separate lines. Stata and R use a hard return.
3. Stata uses `ln()` to compute natural logarithms. R and MATLAB use `log()`.
4. The symbol = is used to define a variable. R prefers <-. Double equality == is used to test equality.

We now illustrate programming files for Stata, R, and MATLAB, which execute a portion of the empirical illustrations from Sections 3.7 and 3.21. For the R and MATLAB code we illustrate using explicit matrix operations. Alternatively, R and MATLAB have built-in functions which implement least squares regression without the need for explicit matrix operations. In R the standard function is `lm`. In MATLAB the standard function is `regress`. The advantage of using explicit matrix operations as shown below is that you know exactly what computations are done and it is easier to go "out of the box" to execute new procedures. The advantage of using built-in functions is that coding is simplified and you are much less likely to make a coding error.

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 mbf12 = (mbf == 1) & (experience == 12)
gen sam = (race == 4) & (marital == 7) & (female == 0)
*        Regressions
reg wage education if mbf12 == 1
reg wage education experience exp2 if sam == 1
*       Leverage and influence
predict leverage, hat
predict e, residual
gen d=e*leverage/(1-leverage)
summarize d if sam ==1
```

R Program File

```
#       Load the data and create subsamples
dat <- read.table("cps09mar.txt")
experience <- dat[,1]-dat[,4]-6
mbf <- (dat[,11]==2)&(dat[,12]<=2)&(dat[,2]==1)&(experience==12)
sam <- (dat[,11]==4)&(dat[,12]==7)&(dat[,2]==0)
dat1 <- dat[mbf,]
dat2 <- dat[sam,]
#       First regression
y <- as.matrix(log(dat1[,5]/(dat1[,6]*dat1[,7])))
x <- cbind(dat1[,4],matrix(1,nrow(dat1),1))
xx <- t(x)%*%x
xy <- t(x)%*%y
beta <- solve(xx,xy)
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))
xx <- t(x)%*%x
xy <- t(x)%*%y
beta <- solve(xx,xy)
print(beta)
#       Create leverage and influence
e <- y-x%*%beta
xxi <- solve(xx)
leverage <- rowSums(x*(x%*%xxi))
r <- e/(1-leverage)
d <- leverage*e/(1-leverage)
print(max(abs(d)))
```

MATLAB Program File

```
% Load the data and create subsamples
dat = load cps09mar.txt;
# An alternative to load the data from an excel file is
# dat = xlsread('cps09mar.xlsx');
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)];
xx = x'*x
xy = x'*y
beta = xx\xy;
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)];
xx = x'*x
xy = x'*y
beta = xx\xy;display(beta);
%       Create leverage and influence
e = y-x*beta;
xxi = inv(xx)
leverage = sum((x.*(x*xxi))')';
d = leverage.*e./(1-leverage);
influence = max(abs(d));
display(influence);
```

3.26 Exercises

Exercise 3.1 Let Y be a random variable with $\mu = \mathbb{E}[Y]$ and $\sigma^2 = \text{var}[Y]$. Define

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

Let $(\widehat{\mu}, \widehat{\sigma}^2)$ be the values such that $\overline{g}_n(\widehat{\mu}, \widehat{\sigma}^2) = 0$ where $\overline{g}_n(m,s) = n^{-1}\sum_{i=1}^n g\left(y_i, m, s\right)$. Show that $\widehat{\mu}$ and $\widehat{\sigma}^2$ are the sample mean and variance.

Exercise 3.2 Consider the OLS regression of the $n \times 1$ vector $\boldsymbol{Y}$ on the $n \times k$ matrix $\boldsymbol{X}$. Consider an alternative set of regressors $\boldsymbol{Z} = \boldsymbol{XC}$, where $\boldsymbol{C}$ is a $k \times k$ non-singular matrix. Thus, each column of $\boldsymbol{Z}$ is a

mixture of some of the columns of $\boldsymbol{X}$. Compare the OLS estimates and residuals from the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$ to the OLS estimates from the regression of $\boldsymbol{Y}$ on $\boldsymbol{Z}$.

Exercise 3.3 Using matrix algebra, show $\boldsymbol{X}'\widehat{\boldsymbol{e}} = 0$.

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

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

Exercise 3.6 Let $\widehat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$. Find the OLS coefficient from a regression of $\widehat{\boldsymbol{Y}}$ on $\boldsymbol{X}$.

Exercise 3.7 Show that if $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2]$ then $\boldsymbol{P}\boldsymbol{X}_1 = \boldsymbol{X}_1$ and $\boldsymbol{M}\boldsymbol{X}_1 = 0$.

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, $n^{-1}\sum_{i=1}^{n}\widehat{Y}_i = \overline{Y}$.

Exercise 3.12 A dummy variable takes on only the values 0 and 1. It is used for categorical variables. 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

$$\boldsymbol{Y} = \mu + \boldsymbol{D}_1\alpha_1 + \boldsymbol{D}_2\alpha_2 + \boldsymbol{e} \tag{3.52}$$

$$\boldsymbol{Y} = \boldsymbol{D}_1\alpha_1 + \boldsymbol{D}_2\alpha_2 + \boldsymbol{e} \tag{3.53}$$

$$\boldsymbol{Y} = \mu + \boldsymbol{D}_1\phi + \boldsymbol{e} \tag{3.54}$$

Can all three equations (3.52), (3.53), and (3.54) be estimated by OLS? Explain if not.

(a) Compare regressions (3.53) and (3.54). Is one more general than the other? Explain the relationship between the parameters in (3.53) and (3.54).

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

Exercise 3.13 Let $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ be defined as in the previous exercise.

(a) In the OLS regression

$$\boldsymbol{Y} = \boldsymbol{D}_1\widehat{\gamma}_1 + \boldsymbol{D}_2\widehat{\gamma}_2 + \widehat{\boldsymbol{u}},$$

show that $\widehat{\gamma}_1$ is the sample mean of the dependent variable among the men of the sample ($\overline{Y}_1$), and that $\widehat{\gamma}_2$ is the sample mean among the women ($\overline{Y}_2$).

(b) Let $\boldsymbol{X}$ ($n \times k$) be an additional matrix of regressors. Describe in words the transformations

$$\boldsymbol{Y}^* = \boldsymbol{Y} - \boldsymbol{D}_1\overline{Y}_1 - \boldsymbol{D}_2\overline{Y}_2$$

$$\boldsymbol{X}^* = \boldsymbol{X} - \boldsymbol{D}_1\overline{X}_1' - \boldsymbol{D}_2\overline{X}_2'$$

where $\overline{X}_1$ and $\overline{X}_2$ are the $k \times 1$ means of the regressors for men and women, respectively.

(c) Compare $\widetilde{\beta}$ from the OLS regression

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

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

$$\boldsymbol{Y} = \boldsymbol{D}_1 \widehat{\alpha}_1 + \boldsymbol{D}_2 \widehat{\alpha}_2 + \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{e}}.$$

Exercise 3.14 Let $\widehat{\beta}_n = \left(\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1}\boldsymbol{X}_n'\boldsymbol{Y}_n$ 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}, X_{n+1}) becomes available. Prove that the OLS estimate computed using this additional observation is

$$\widehat{\beta}_{n+1} = \widehat{\beta}_n + \frac{1}{1 + X_{n+1}'\left(\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1} X_{n+1}} \left(\boldsymbol{X}_n'\boldsymbol{X}_n\right)^{-1} X_{n+1}\left(Y_{n+1} - X_{n+1}'\widehat{\beta}_n\right).$$

Exercise 3.15 Prove that R^2 is the square of the sample correlation between $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$.

Exercise 3.16 Consider two least squares regressions

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

and

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

Let R_1^2 and R_2^2 be the R-squared from the two regressions. Show that $R_2^2 \geq R_1^2$. Is there a case (explain) when there is equality $R_2^2 = R_1^2$?

Exercise 3.17 For $\widetilde{\sigma}^2$ defined in (3.46), show that $\widetilde{\sigma}^2 \geq \widehat{\sigma}^2$. Is equality possible?

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

Exercise 3.19 For the intercept-only model $Y_i = \beta + e_i$, show that the leave-one-out prediction error is

$$\widetilde{e}_i = \left(\frac{n}{n-1}\right)\left(Y_i - \overline{Y}\right).$$

Exercise 3.20 Define the leave-one-out estimator of σ^2,

$$\widehat{\sigma}^2_{(-i)} = \frac{1}{n-1}\sum_{j\neq i}\left(Y_j - X_j'\widehat{\beta}_{(-i)}\right)^2.$$

This is the estimator obtained from the sample with observation i omitted. Show that

$$\widehat{\sigma}^2_{(-i)} = \frac{n}{n-1}\widehat{\sigma}^2 - \frac{\widehat{e}_i^2}{(n-1)\left(1-h_{ii}\right)}.$$

Exercise 3.21 Consider the least squares regression estimators

$$Y_i = X_{1i}\widehat{\beta}_1 + X_{2i}\widehat{\beta}_2 + \widehat{e}_i$$

and the "one regressor at a time" regression estimators

$$Y_i = X_{1i}\widetilde{\beta}_1 + \widetilde{e}_{1i}, \qquad Y_i = X_{2i}\widetilde{\beta}_2 + \widetilde{e}_{2i}$$

Under what condition does $\widetilde{\beta}_1 = \widehat{\beta}_1$ and $\widetilde{\beta}_2 = \widehat{\beta}_2$?

Exercise 3.22 You estimate a least squares regression

$$Y_i = X_{1i}'\widetilde{\beta}_1 + \widetilde{u}_i$$

and then regress the residuals on another set of regressors

$$\widetilde{u}_i = X_{2i}'\widetilde{\beta}_2 + \widetilde{e}_i$$

Does this second regression give you the same estimated coefficients as from estimation of a least squares regression on both set of regressors?

$$Y_i = X_{1i}'\widehat{\beta}_1 + X_{2i}'\widehat{\beta}_2 + \widehat{e}_i$$

In other words, is it true that $\widetilde{\beta}_2 = \widehat{\beta}_2$? Explain your reasoning.

Exercise 3.23 The data matrix is $(\boldsymbol{Y}, \boldsymbol{X})$ with $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$, and consider the transformed regressor matrix $\boldsymbol{Z} = [\boldsymbol{X}_1, \boldsymbol{X}_2 - \boldsymbol{X}_1]$. Suppose you do a least squares regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$, and a least squares regression of $\boldsymbol{Y}$ on $\boldsymbol{Z}$. Let $\widehat{\sigma}^2$ and $\widetilde{\sigma}^2$ denote the residual variance estimates from the two regressions. Give a formula relating $\widehat{\sigma}^2$ and $\widetilde{\sigma}^2$? (Explain your reasoning.)

Exercise 3.24 Use the `cps09mar` data set described in Section 3.22 and available on the textbook website. Take the sub-sample used for equation (3.49) (see Section 3.25) for data construction)

(a) Estimate equation (3.49) and compute the equation R^2 and sum of squared errors.

(b) Re-estimate the slope on education using the residual regression approach. Regress log(wage) on experience and its square, regress education on experience and its square, and the residuals on the residuals. Report the estimates from this final regression, along with the equation R^2 and sum of squared errors. Does the slope coefficient equal the value in (3.49)? Explain.

(c) Are the R^2 and sum-of-squared errors from parts (a) and (b) equal? Explain.

Exercise 3.25 Estimate equation (3.49) as in part (a) of the previous question. Let $\widehat{e}_i$ be the OLS residual, $\widehat{Y}_i$ the predicted value from the regression, X_{1i} be education and X_{2i} be experience. Numerically calculate the following:

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

(b) $\sum_{i=1}^n X_{1i}\widehat{e}_i$

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

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

(e) $\sum_{i=1}^n X_{2i}^2\widehat{e}_i$

(f) $\sum_{i=1}^n \widehat{Y}_i\widehat{e}_i$

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

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

Exercise 3.26 Use the `cps09mar` data set.

(a) Estimate a log wage regression for the subsample of white male Hispanics. In addition to education, experience, and its square, include a set of binary variables for regions and marital status. For regions, create dummy variables for Northeast, South, and West so that Midwest is the excluded group. For marital status, create variables for married, widowed or divorced, and separated, so that single (never married) is the excluded group.

(b) Repeat using a different econometric package. Compare your results. Do they agree?

Chapter 4

Least Squares Regression

4.1 Introduction

In this chapter we investigate some finite-sample properties of the least squares estimator in the linear regression model. In particular we calculate its finite-sample expectation and covariance matrix and propose standard errors for the coefficient estimators.

4.2 Random Sampling

Assumption 3.1 specified that the observations have identical distributions. To derive the finite-sample properties of the estimators we will need to additionally specify the dependence structure across the observations.

The simplest context is when the observations are mutually independent in which case we say that they are **independent and identically distributed** or **i.i.d.** It is also common to describe i.i.d. observations as a **random sample**. Traditionally, random sampling has been the default assumption in cross-section (e.g. survey) contexts. It is quite convenient as i.i.d. sampling leads to straightforward expressions for estimation variance. The assumption seems appropriate (meaning that it should be approximately valid) when samples are small and relatively dispersed. That is, if you randomly sample 1000 people from a large country such as the United States it seems reasonable to model their responses as mutually independent.

Assumption 4.1 The random variables $\{(Y_1, X_1), ..., (Y_i, X_i), ..., (Y_n, X_n)\}$ are independent and identically distributed.

For most of this chapter we will use Assumption 4.1 to derive properties of the OLS estimator.

Assumption 4.1 means that if you take any two individuals $i \neq j$ in a sample, the values (Y_i, X_i) are independent of the values (Y_j, X_j) yet have the same distribution. Independence means that the decisions and choices of individual i do not affect the decisions of individual j and conversely.

This assumption may be violated if individuals in the sample are connected in some way, for example if they are neighbors, members of the same village, classmates at a school, or even firms within a specific industry. In this case it seems plausible that decisions may be inter-connected and thus mutually dependent rather than independent. Allowing for such interactions complicates inference and requires specialized treatment. A currently popular approach which allows for mutual dependence is known as

clustered dependence which assumes that that observations are grouped into "clusters" (for example, schools). We will discuss clustering in more detail in Section 4.21.

4.3 Sample Mean

We start with the simplest setting of the intercept-only model

$$\begin{aligned} Y &= \mu + e \\ \mathbb{E}[e] &= 0. \end{aligned}$$

which is equivalent to the regression model with $k = 1$ and $X = 1$. In the intercept model $\mu = \mathbb{E}[Y]$ is the expectation of Y. (See Exercise 2.15.) The least squares estimator $\widehat{\mu} = \overline{Y}$ equals the sample mean as shown in equation (3.8).

We now calculate the expectation 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 the least squares estimator (the sample mean) equals the projection coefficient (the population expectation). An estimator with the property that its expectation equals the parameter it is estimating is called **unbiased**.

Definition 4.1 An estimator $\widehat{\theta}$ for θ is **unbiased** if $\mathbb{E}\left[\widehat{\theta}\right] = \theta$.

We next calculate the variance of the estimator $\overline{Y}$ under Assumption 4.1. Making the substitution $Y_i = \mu + e_i$ we find

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

Then

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

The second-to-last inequality is because $\mathbb{E}\left[e_i e_j\right] = \sigma^2$ for $i = j$ yet $\mathbb{E}\left[e_i e_j\right] = 0$ for $i \neq j$ due to independence.

We have shown that $\operatorname{var}\left[\overline{Y}\right] = \frac{1}{n}\sigma^2$. This is the familiar formula for the variance of the sample mean.

4.4 Linear Regression Model

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

Assumption 4.2 Linear Regression Model
The variables (Y, X) satisfy the linear regression equation

$$Y = X'\beta + e \tag{4.1}$$

$$\mathbb{E}[e \mid X] = 0. \tag{4.2}$$

The variables have finite second moments

$$\mathbb{E}\left[Y^2\right] < \infty,$$

$$\mathbb{E}\|X\|^2 < \infty,$$

and an invertible design matrix

$$\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right] > 0.$$

We will consider both the general case of heteroskedastic regression where the conditional variance $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2(X)$ 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 Homoskedastic Linear Regression Model
In addition to Assumption 4.2

$$\mathbb{E}\left[e^2 \mid X\right] = \sigma^2(X) = \sigma^2 \tag{4.3}$$

is independent of X.

4.5 Expectation 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)

$$\mathbb{E}[Y_i \mid X_1, ..., X_n] = \mathbb{E}[Y_i \mid X_i] = X_i'\beta. \tag{4.4}$$

The first equality states that the conditional expectation of Y_i given $\{X_1, ..., X_n\}$ only depends on X_i because the observations are independent across i. The second equality is the assumption of a linear conditional expectation.

Using definition (3.11), the conditioning theorem (Theorem 2.3), the linearity of expectations, (4.4), and properties of the matrix inverse,

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta} \mid X_1, ..., X_n\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right)\middle| X_1, ..., X_n\right] \\
&= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \mathbb{E}\left[\left(\sum_{i=1}^{n} X_i Y_i\right)\middle| X_1, ..., X_n\right] \\
&= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} \mathbb{E}\left[X_i Y_i \mid X_1, ..., X_n\right] \\
&= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i \mathbb{E}\left[Y_i \mid X_i\right] \\
&= \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i X_i' \beta \\
&= \beta.
\end{aligned}$$

Now let's show the same result using matrix notation. (4.4) implies

$$\mathbb{E}\left[\boldsymbol{Y} \mid \boldsymbol{X}\right] = \begin{pmatrix} \vdots \\ \mathbb{E}\left[Y_i \mid \boldsymbol{X}\right] \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ X_i'\beta \\ \vdots \end{pmatrix} = \boldsymbol{X}\beta. \tag{4.5}$$

Similarly

$$\mathbb{E}\left[\boldsymbol{e} \mid \boldsymbol{X}\right] = \begin{pmatrix} \vdots \\ \mathbb{E}\left[e_i \mid \boldsymbol{X}\right] \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathbb{E}\left[e_i \mid X_i\right] \\ \vdots \end{pmatrix} = 0.$$

Using $\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right)$, the conditioning theorem, the linearity of expectations, (4.5), and the properties of the matrix inverse,

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

At the risk of belaboring the derivation, another way to calculate the same result is as follows. Insert $\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}$ into the formula for $\widehat{\beta}$ to obtain

$$\begin{aligned}
\widehat{\beta} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\left(\boldsymbol{X}\beta + \boldsymbol{e}\right)\right) \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\beta + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{e}\right) \\
&= \beta + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}.
\end{aligned} \tag{4.6}$$

This is a useful linear decomposition of the estimator $\widehat{\beta}$ into the true parameter β and the stochastic component $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}$. Once again, we can calculate that

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

Regardless of the method we have shown that $\mathbb{E}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \beta$. We have shown the following theorem.

Theorem 4.1 Expectation of Least Squares Estimator
In the linear regression model (Assumption 4.2) with i.i.d. sampling (Assumption 4.1)

$$\mathbb{E}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \beta. \tag{4.7}$$

Equation (4.7) says that the estimator $\widehat{\beta}$ is unbiased for β, conditional on $\boldsymbol{X}$. This means that the conditional distribution of $\widehat{\beta}$ is centered at β. By "conditional on $\boldsymbol{X}$" this means that the distribution is unbiased for any realization of the regressor matrix $\boldsymbol{X}$. In conditional models we simply refer to this as saying "$\widehat{\beta}$ is unbiased for β".

It is worth mentioning that Theorem 4.1, and all finite sample results in this chapter, make the implicit assumption that $\boldsymbol{X}'\boldsymbol{X}$ is full rank with probability one.

4.6 Variance of Least Squares Estimator

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

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

$$\operatorname{var}[Z] = \mathbb{E}\left[(Z - \mathbb{E}[Z])(Z - \mathbb{E}[Z])'\right] = \mathbb{E}\left[ZZ'\right] - (\mathbb{E}[Z])(\mathbb{E}[Z])'$$

and for any pair (Z, X) define the conditional covariance matrix

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

We define $\boldsymbol{V}_{\widehat{\beta}} \stackrel{\text{def}}{=} \operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right]$ as the conditional covariance matrix of the regression coefficient estimators. We now derive its form.

The conditional covariance matrix of the $n \times 1$ regression error $\boldsymbol{e}$ is the $n \times n$ matrix

$$\operatorname{var}[\boldsymbol{e} \mid \boldsymbol{X}] = \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right] \stackrel{\text{def}}{=} \boldsymbol{D}.$$

The i^{th} diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}\left[e_i^2 \mid \boldsymbol{X}\right] = \mathbb{E}\left[e_i^2 \mid X_i\right] = \sigma_i^2$$

while the ij^{th} off-diagonal element of $\boldsymbol{D}$ is

$$\mathbb{E}\left[e_i e_j \mid \boldsymbol{X}\right] = \mathbb{E}\left(e_i \mid X_i\right)\mathbb{E}\left[e_j \mid X_j\right] = 0$$

where the first equality uses independence of the observations (Assumption 4.1) and the second is (4.2). Thus $\boldsymbol{D}$ is a diagonal matrix with i^{th} diagonal element σ_i^2:

$$\boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}. \tag{4.8}$$

In the special case of the linear homoskedastic regression model (4.3), then $\mathbb{E}\left[e_i^2 \mid X_i\right] = \sigma_i^2 = \sigma^2$ 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 $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{X})$,

$$\operatorname{var}\left[\boldsymbol{A}'\boldsymbol{Y} \mid \boldsymbol{X}\right] = \operatorname{var}\left[\boldsymbol{A}'\boldsymbol{e} \mid \boldsymbol{X}\right] = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A}. \tag{4.9}$$

In particular, we can write $\widehat{\beta} = \boldsymbol{A}'\boldsymbol{Y}$ where $\boldsymbol{A} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and thus

$$\boldsymbol{V}_{\widehat{\beta}} = \operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

It is useful to note that

$$\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X} = \sum_{i=1}^{n} X_i X_i' \sigma_i^2,$$

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

Theorem 4.2 Variance of Least Squares Estimator
In the linear regression model (Assumption 4.2) with i.i.d. sampling (Assumption 4.1)

$$\boldsymbol{V}_{\widehat{\beta}} = \operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{4.10}$$

where $\boldsymbol{D}$ is defined in (4.8). If in addition the error is homoskedastic (Assumption 4.3) then (4.10) simplifies to $\boldsymbol{V}_{\widehat{\beta}} = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$.

4.7 Unconditional Moments

The previous sections derived the form of the conditional expectation and variance of the least squares estimator where we conditioned on the regressor matrix $\boldsymbol{X}$. What about the unconditional expectation and variance?

Indeed, it is not obvious if $\widehat{\beta}$ has a finite expectation or variance. Take the case of a single dummy variable regressor D_i with no intercept. Assume $\mathbb{P}[D_i = 1] = p < 1$. Then

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} D_i Y_i}{\sum_{i=1}^{n} D_i}$$

is well defined if $\sum_{i=1}^{n} D_i > 0$. However, $\mathbb{P}\left[\sum_{i=1}^{n} D_i = 0\right] = \left(1-p\right)^n > 0$. This means that with positive (but small) probability $\widehat{\beta}$ does not exist. Consequently $\widehat{\beta}$ has no finite moments! We ignore this complication in practice but it does pose a conundrum for theory. This existence problem arises whenever there are discrete regressors.

This dilemma is avoided when the regressors have continuous distributions. A clean statement was obtained by Kinal (1980) under the assumption of normal regressors and errors.

Theorem 4.3 Kinal (1980)
In the linear regression model with i.i.d. sampling, if in addition (X,e) have a joint normal distribution, then for any r, $\mathbb{E}\left\|\widehat{\beta}\right\|^r < \infty$ if and only if $r < n-k+1$.

This shows that when the errors and regressors are normally distributed that the least squares estimator possesses all moments up to $n-k$ which includes all moments of practical interest. The normality assumption is not critical for this result. What is key is the assumption that the regressors are continuously distributed.

The law of iterated expectations (Theorem 2.1) combined with Theorems 4.1 and 4.3 allow us to deduce that the least squares estimator is unconditionally unbiased. Under the normality assumption Theorem 4.3 allows us to apply the law of iterated expectations, and thus using Theorems 4.1 we deduce that if $n > k$

$$\mathbb{E}\left[\widehat{\beta}\right] = \mathbb{E}\left[\mathbb{E}\left[\widehat{\beta} \mid \boldsymbol{X}\right]\right] = \beta.$$

Hence $\widehat{\beta}$ is unconditionally unbiased as asserted.

Furthermore, if $n-k > 1$ then $\mathbb{E}\left\|\widehat{\beta}\right\|^2 < \infty$ and $\widehat{\beta}$ has a finite unconditional variance. Using Theorem 2.8 we can calculate explicitly that

$$\operatorname{var}\left[\widehat{\beta}\right] = \mathbb{E}\left[\operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right]\right] + \operatorname{var}\left[\mathbb{E}\left[\widehat{\beta} \mid \boldsymbol{X}\right]\right] = \mathbb{E}\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]$$

the second equality because $\mathbb{E}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \beta$ has zero variance. In the homoskedastic case this simplifies to

$$\operatorname{var}\left[\widehat{\beta}\right] = \sigma^2 \mathbb{E}\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right].$$

In both cases the expectation cannot pass through the matrix inverse because this is a nonlinear function. Thus there is not a simple expression for the unconditional variance, other than stating that is it the expectation of the conditional variance.

4.8 Gauss-Markov Theorem

The Gauss-Markov Theorem is one of the most celebrated results in econometric theory. It provides a classical justification for the least squares estimator, showing that it is lowest variance among unbiased estimators.

Write the homoskedastic linear regression model in vector format as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e} \tag{4.11}$$

$$\mathbb{E}\left[\boldsymbol{e} \mid \boldsymbol{X}\right] = 0 \tag{4.12}$$

$$\operatorname{var}\left[\boldsymbol{e} \mid \boldsymbol{X}\right] = \boldsymbol{I}_n \sigma^2. \tag{4.13}$$

In this model we know that the least squares estimator is unbiased for β and has covariance matrix $\sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. The question raised in this section is if there exists an alternative unbiased estimator $\widetilde{\beta}$ which has a smaller covariance matrix.

The following version of the theorem is due to B. E. Hansen (2021).

Theorem 4.4 Gauss-Markov
Take the homoskedastic linear regression model (4.11)-(4.13). If $\widetilde{\beta}$ is an unbiased estimator of β then

$$\operatorname{var}\left[\widetilde{\beta} \mid \boldsymbol{X}\right] \geq \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

Theorem 4.4 provides a lower bound on the covariance matrix of unbiased estimators under the assumption of homoskedasticity. It says that no unbiased estimator can have a variance matrix smaller (in the positive definite sense) than $\sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. Since the variance of the OLS estimator is exactly equal to this bound this means that no unbiased estimator has a lower variance than OLS. Consequently we describe OLS as **efficient** in the class of unbiased estimators.

This earliest version of Theorem 4.4 was articulated by Carl Friedrich Gauss in 1823. Andreĭ Andreevich Markov provided a textbook treatment of the theorem in 1912, and clarified the central role of unbiasedness, which Gauss had only assumed implicitly.

Their versions of the Theorem restricted attention to **linear** estimators of β, which are estimators that can be written as $\widetilde{\beta} = \boldsymbol{A}'\boldsymbol{Y}$, where $\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{X})$ is an $m \times n$ function of the regressors $\boldsymbol{X}$. Linearity in this context means "linear in $\boldsymbol{Y}$". This restriction simplifies variance calculations, but greatly limits the class of estimators.This classical version of the Theorem gave rise to the description of OLS as the **best linear unbiased estimator (BLUE)**. However, Theorem 4.4 as stated above shows that OLS is the **best unbiased estimator (BUE)**.

The derivation of the Gauss-Markov Theorem under the restriction to linear estimators is straightforward, so we now provide this demonstration. For $\widetilde{\beta} = \boldsymbol{A}'\boldsymbol{Y}$ we have

$$\mathbb{E}\left[\widetilde{\beta} \mid \boldsymbol{X}\right] = \boldsymbol{A}'\mathbb{E}\left[\boldsymbol{Y} \mid \boldsymbol{X}\right] = \boldsymbol{A}'\boldsymbol{X}\beta,$$

the second equality because $\mathbb{E}\left[\boldsymbol{Y} \mid \boldsymbol{X}\right] = \boldsymbol{X}\beta$. Then $\widetilde{\beta}$ is unbiased for all β if (and only if) $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{I}_k$. Furthermore, we saw in (4.9) that

$$\operatorname{var}\left[\widetilde{\beta} \mid \boldsymbol{X}\right] = \operatorname{var}\left[\boldsymbol{A}'\boldsymbol{Y} \mid \boldsymbol{X}\right] = \boldsymbol{A}'\boldsymbol{D}\boldsymbol{A} = \boldsymbol{A}'\boldsymbol{A}\sigma^2$$

the last equality using the homoskedasticity assumption (4.13). To establish the Theorem we need to show that for any such matrix $\boldsymbol{A}$,

$$\boldsymbol{A}'\boldsymbol{A} \geq \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.14}$$

Set $\boldsymbol{C} = \boldsymbol{A} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. Note that $\boldsymbol{X}'\boldsymbol{C} = 0$. We calculate that

$$\begin{aligned}
\boldsymbol{A}'\boldsymbol{A} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} &= \left(\boldsymbol{C} + \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)'\left(\boldsymbol{C} + \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{C}'\boldsymbol{C} + \boldsymbol{C}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{C} \\
&\quad + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \boldsymbol{C}'\boldsymbol{C} \geq 0.
\end{aligned}$$

The final inequality states that the matrix $\boldsymbol{C}'\boldsymbol{C}$ is positive semi-definite which is a property of quadratic forms (see Appendix A.10). We have shown (4.14) as required.

The above derivation imposed the restriction that the estimator $\widetilde{\beta}$ is linear in $\boldsymbol{Y}$. The proof of Theorem 4.4 in the general case is considerably more advanced. Here, we provide a simplified sketch of the argument for interested readers, with a complete proof in Section 4.24.

For simplicity, treat the regressors X as fixed, and suppose that Y has a density $f(y)$ with bounded support $\mathscr{Y}$. Without loss of generality assume that the true coefficient equals $\beta_0 = 0$.

Since Y has bounded support $\mathscr{Y}$ there is a set $B \subset \mathbb{R}^m$ such that $\left|yX'\beta/\sigma^2\right| < 1$ for all $\beta \in B$ and $y \in \mathscr{Y}$. For such values of β, define the auxiliary density function

$$f_\beta(y) = f(y)\left(1 + yX'\beta/\sigma^2\right). \tag{4.15}$$

Under the assumptions, $0 \le f_\beta(y) \le 2f(y)$, $f_\beta(y)$ has support $\mathscr{Y}$, and $\int_{\mathscr{Y}} f_\beta(y)dy = 1$. To see the later, observe that $\int_{\mathscr{Y}} yf(y)dy = X'\beta_0 = 0$ under the normalization $\beta_0 = 0$, and thus

$$\int_{\mathscr{Y}} f_\beta(y)dy = \int_{\mathscr{Y}} f(y)dy + \int_{\mathscr{Y}} f(y)ydyX'\beta/\sigma^2 = 1$$

because $\int_{\mathscr{Y}} f(y)dy = 1$. Thus f_β is a parametric family of density functions. Evaluated at β_0 we see that $f_0 = f$, which means that f_β is a correctly-specified parametric family with true parameter value $\beta_0 = 0$.

To illustrate, take the case $X = 1$. Figure 4.1 displays an example density $f(y) = (3/4)(1-y^2)$ on $[-1,1]$ with auxiliary density $f_\beta(y) = f(y)\left(1+y\right)$. We can see how the auxiliary density is a tilted version of the original density $f(y)$.

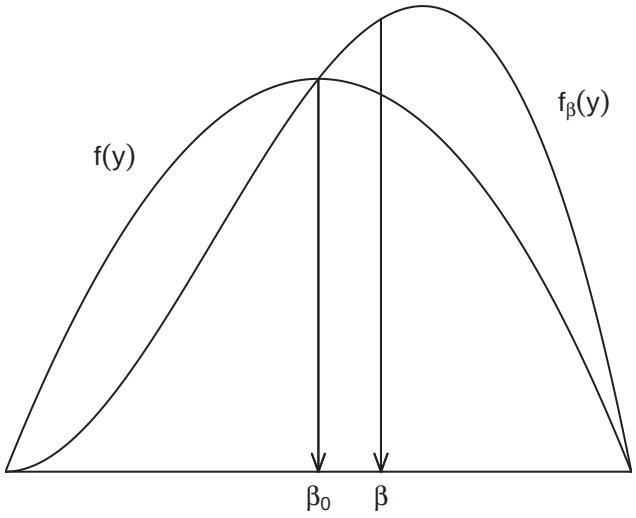


Figure 4.1: Original and Auxiliary Density

Let $\mathbb{E}_\beta$ denote expectation with respect to the auxiliary distribution. Since $\int_{\mathscr{Y}} yf(y)dy = 0$ and $\int_{\mathscr{Y}} y^2 f(y)dy = \sigma^2$, we find

$$\mathbb{E}_\beta[Y] = \int_{\mathscr{Y}} yf_\beta(y)dy = \int_{\mathscr{Y}} yf(y)dy + \int_{\mathscr{Y}} y^2 f(y)dyX'\beta/\sigma^2 = X'\beta.$$

This shows that f_β is a regression model with regression coefficient β.

In Figure 4.1, the means of the two densities are indicated by the arrows to the x-axis. In this example we can see how the auxiliary density has a larger expected value, because the density has been tilted to the right.

The parametric family f_β over $\beta \in B$ has the following properties: its expectation is $X'\beta$, its variance is finite, the true value β_0 lies in the interior of B, and the support of the distribution does not depend on β.

The likelihood score of the auxiliary density function for an observation, using the fact that $Y_i = e_i$, is

$$S_i = \frac{\partial}{\partial \beta}\left(\log f_\beta(Y_i)\right)\Big|_{\beta=0} = \frac{\partial}{\partial \beta}\left(\log f(e_i) + \log\left(1 + e_i X_i'\beta/\sigma^2\right)\right)\Big|_{\beta=0} = X_i e_i/\sigma^2. \tag{4.16}$$

Therefore the information matrix is

$$\mathscr{I} = \sum_{i=1}^{n} \mathbb{E}\left[S_i S_i'\right] = \sum_{i=1}^{n} X_i X_i' \mathbb{E}\left[e_i^2\right]/\sigma^4 = \left(\boldsymbol{X}'\boldsymbol{X}\right)/\sigma^2.$$

By assumption, $\widetilde{\beta}$ is unbiased. The Cramér-Rao lower bound states that

$$\operatorname{var}\left[\widetilde{\beta}\right] \geq \mathscr{I}^{-1} = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.$$

This is the variance lower bound, completing the proof of Theorem 4.4.

The above argument is rather tricky. At its core is the observation that the model f_β is a submodel of the set of all linear regression models. The Cramér-Rao bound over any regular parametric submodel is a lower bound on the variance of any unbiased estimator. This means that the Cramér-Rao bound over f_β is a lower bound for unbiased estimation of the regression coefficient. The model f_β was selected judiciously so that its Cramér-Rao bound equals the variance of the least squares estimator, and this is sufficient to establish the bound.

4.9 Generalized Least Squares

Take the linear regression model in matrix format

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}. \tag{4.17}$$

Consider a generalized situation where the observation errors are possibly correlated and/or heteroskedastic. Specifically, suppose that

$$\mathbb{E}\left[\boldsymbol{e} \mid \boldsymbol{X}\right] = 0 \tag{4.18}$$

$$\operatorname{var}\left[\boldsymbol{e} \mid \boldsymbol{X}\right] = \Sigma\sigma^2 \tag{4.19}$$

for some $n \times n$ matrix $\Sigma > 0$, possibly a function of $\boldsymbol{X}$, and some scalar σ^2. This includes the independent sampling framework where Σ is diagonal but allows for non-diagonal covariance matrices as well. As a scaled covariance matrix, Σ is necessarily symmetric and positive semi-definite.

Under these assumptions, by arguments similar to the previous sections we can calculate the expectation and variance of the OLS estimator:

$$\mathbb{E}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \beta \tag{4.20}$$

$$\operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\Sigma\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{4.21}$$

(see Exercise 4.5).

Aitken (1935) established a generalization of the Gauss-Markov Theorem. The following statement is due to B. E. Hansen (2021).

Theorem 4.5 Take the linear regression model (4.17)-(4.19). If $\widetilde{\beta}$ is an unbiased estimator of β then

$$\operatorname{var}\left[\widetilde{\beta} \mid \boldsymbol{X}\right] \geq \sigma^2 \left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}.$$

We defer the proof to Section 4.24. See also Exercise 4.6.

Theorem 4.5 provides a lower bound on the covariance matrix of unbiased estimators. Theorem 4.4 was the special case $\Sigma = \boldsymbol{I}_n$.

When Σ is known, Aitken (1935) constructed an estimator which achieves the lower bound in Theorem 4.5. Take the linear model (4.17) and pre-multiply by $\Sigma^{-1/2}$. This produces the equation $\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{X}}\beta + \widetilde{\boldsymbol{e}}$ where $\widetilde{\boldsymbol{Y}} = \Sigma^{-1/2}\boldsymbol{Y}$, $\widetilde{\boldsymbol{X}} = \Sigma^{-1/2}\boldsymbol{X}$, and $\widetilde{\boldsymbol{e}} = \Sigma^{-1/2}\boldsymbol{e}$. Consider OLS estimation of β in this equation.

$$\begin{aligned}
\widetilde{\beta}_{\text{gls}} &= \left(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}}\right)^{-1}\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{Y}} \\
&= \left(\left(\Sigma^{-1/2}\boldsymbol{X}\right)'\left(\Sigma^{-1/2}\boldsymbol{X}\right)\right)^{-1}\left(\Sigma^{-1/2}\boldsymbol{X}\right)'\left(\Sigma^{-1/2}\boldsymbol{Y}\right) \\
&= \left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\Sigma^{-1}\boldsymbol{Y}. \qquad (4.22)
\end{aligned}$$

This is called the **Generalized Least Squares** (GLS) estimator of β.

You can calculate that

$$\mathbb{E}\left[\widetilde{\beta}_{\text{gls}} \mid \boldsymbol{X}\right] = \beta \qquad (4.23)$$

$$\operatorname{var}\left[\widetilde{\beta}_{\text{gls}} \mid \boldsymbol{X}\right] = \sigma^2\left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}. \qquad (4.24)$$

This shows that the GLS estimator is unbiased and has a covariance matrix which equals the lower bound from Theorem 4.5. This shows that the lower bound is sharp. GLS is thus efficient in the class of unbiased estimators.

In the linear regression model with independent observations and known conditional variances, so that $\Sigma = \boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right)$, the GLS estimator takes the form

$$\begin{aligned}
\widetilde{\beta}_{\text{gls}} &= \left(\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}^{-1}\boldsymbol{Y} \\
&= \left(\sum_{i=1}^{n}\sigma_i^{-2}X_iX_i'\right)^{-1}\left(\sum_{i=1}^{n}\sigma_i^{-2}X_iY_i\right).
\end{aligned}$$

The assumption $\Sigma > 0$ in this case reduces to $\sigma_i^2 > 0$ for $i = 1, ...n$.

In most settings the matrix Σ is unknown so the GLS estimator is not feasible. However, the form of the GLS estimator motivates feasible versions, effectively by replacing Σ with a suitable estimator.

4.10 Residuals

What are some properties of the residuals $\widehat{e}_i = Y_i - X_i'\widehat{\beta}$ and prediction errors $\widetilde{e}_i = Y_i - X_i'\widehat{\beta}_{(-i)}$ in the context of the linear regression model?

Recall from (3.24) that we can write the residuals in vector notation as $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ is the orthogonal projection matrix. Using the properties of conditional expectation

$$\mathbb{E}\left[\widehat{\boldsymbol{e}} \mid \boldsymbol{X}\right] = \mathbb{E}\left[\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}\right] = \boldsymbol{M}\mathbb{E}\left[\boldsymbol{e} \mid \boldsymbol{X}\right] = 0$$

and

$$\operatorname{var}\left[\widehat{\boldsymbol{e}} \mid \boldsymbol{X}\right] = \operatorname{var}\left[\boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X}\right] = \boldsymbol{M}\operatorname{var}\left[\boldsymbol{e} \mid \boldsymbol{X}\right]\boldsymbol{M} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{M} \qquad (4.25)$$

where $\boldsymbol{D}$ is defined in (4.8).

We can simplify this expression under the assumption of conditional homoskedasticity

$$\mathbb{E}\left[e^2 \mid X\right] = \sigma^2.$$

In this case (4.25) simplifies to

$$\operatorname{var}[\widehat{\boldsymbol{e}} \mid \boldsymbol{X}] = \boldsymbol{M}\sigma^2. \tag{4.26}$$

In particular, for a single observation i we can find the variance of $\widehat{e}_i$ by taking the i^{th} diagonal element of (4.26). Since the i^{th} diagonal element of $\boldsymbol{M}$ is $1 - h_{ii}$ as defined in (3.40) we obtain

$$\operatorname{var}[\widehat{e}_i \mid \boldsymbol{X}] = \mathbb{E}\left[\widehat{e}_i^2 \mid \boldsymbol{X}\right] = (1 - h_{ii})\,\sigma^2. \tag{4.27}$$

As this variance is a function of h_{ii} and hence X_i the residuals $\widehat{e}_i$ are heteroskedastic even if the errors e_i are homoskedastic. Notice as well that (4.27) implies $\widehat{e}_i^2$ is a biased estimator of σ^2.

Similarly, recall from (3.45) that the prediction errors $\widetilde{e}_i = (1 - h_{ii})^{-1}\,\widehat{e}_i$ can be written in vector notation as $\widetilde{\boldsymbol{e}} = \boldsymbol{M}^*\widehat{\boldsymbol{e}}$ where $\boldsymbol{M}^*$ is a diagonal matrix with i^{th} diagonal element $(1 - h_{ii})^{-1}$. Thus $\widetilde{\boldsymbol{e}} = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{e}$. We can calculate that

$$\mathbb{E}[\widetilde{\boldsymbol{e}} \mid \boldsymbol{X}] = \boldsymbol{M}^*\boldsymbol{M}\mathbb{E}[\boldsymbol{e} \mid \boldsymbol{X}] = 0$$

and

$$\operatorname{var}[\widetilde{\boldsymbol{e}} \mid \boldsymbol{X}] = \boldsymbol{M}^*\boldsymbol{M}\operatorname{var}[\boldsymbol{e} \mid \boldsymbol{X}]\,\boldsymbol{M}\boldsymbol{M}^* = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}\boldsymbol{M}^*$$

which simplifies under homoskedasticity to

$$\operatorname{var}[\widetilde{\boldsymbol{e}} \mid \boldsymbol{X}] = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}\boldsymbol{M}^*\sigma^2 = \boldsymbol{M}^*\boldsymbol{M}\boldsymbol{M}^*\sigma^2.$$

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

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

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

$$\overline{e}_i = (1 - h_{ii})^{-1/2}\,\widehat{e}_i, \tag{4.28}$$

and in vector notation

$$\overline{\boldsymbol{e}} = \left(\overline{e}_1, ..., \overline{e}_n\right)' = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{e}. \tag{4.29}$$

From the above calculations, under homoskedasticity,

$$\operatorname{var}\left[\overline{\boldsymbol{e}} \mid \boldsymbol{X}\right] = \boldsymbol{M}^{*1/2}\boldsymbol{M}\boldsymbol{M}^{*1/2}\sigma^2$$

and

$$\operatorname{var}\left[\overline{e}_i \mid \boldsymbol{X}\right] = \mathbb{E}\left[\overline{e}_i^2 \mid \boldsymbol{X}\right] = \sigma^2$$

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

4.11 Estimation of Error Variance

The error variance $\sigma^2 = \mathbb{E}\left[e^2\right]$ can be a parameter of interest even in a heteroskedastic regression or a projection model. σ^2 measures the variation in the "unexplained" part of the regression. Its method of moments estimator (MME) is the sample average of the squared residuals:

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

In the linear regression model we can calculate the expectation of $\widehat{\sigma}^2$. From (3.28) and the properties of the trace operator observe that

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

Then

$$\begin{aligned}
\mathbb{E}\left[\widehat{\sigma}^2 \mid \boldsymbol{X}\right] &= \frac{1}{n}\operatorname{tr}\left(\mathbb{E}\left[\boldsymbol{M}\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right]\right) \\
&= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right]\right) \\
&= \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\boldsymbol{D}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(1 - h_{ii}\right)\sigma_i^2.
\end{aligned} \tag{4.30}$$

The final equality holds because the trace is the sum of the diagonal elements of $\boldsymbol{M}\boldsymbol{D}$, and because $\boldsymbol{D}$ is diagonal the diagonal elements of $\boldsymbol{M}\boldsymbol{D}$ are the product of the diagonal elements of $\boldsymbol{M}$ and $\boldsymbol{D}$ which are $1 - h_{ii}$ and σ_i^2, respectively.

Adding the assumption of conditional homoskedasticity $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$ so that $\boldsymbol{D} = \boldsymbol{I}_n\sigma^2$, then (4.30) simplifies to

$$\mathbb{E}\left[\widehat{\sigma}^2 \mid \boldsymbol{X}\right] = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}\sigma^2\right) = \sigma^2\left(\frac{n-k}{n}\right)$$

the final equality by (3.22). This calculation shows that $\widehat{\sigma}^2$ is biased towards zero. The order of the bias depends on k/n, the ratio of the number of estimated coefficients to the sample size.

Another way to see this is to use (4.27). Note that

$$\mathbb{E}\left[\widehat{\sigma}^2 \mid \boldsymbol{X}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\widehat{e}_i^2 \mid \boldsymbol{X}\right] = \frac{1}{n}\sum_{i=1}^{n}\left(1 - h_{ii}\right)\sigma^2 = \left(\frac{n-k}{n}\right)\sigma^2$$

the last equality using Theorem 3.6.

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

$$s^2 = \frac{1}{n-k}\sum_{i=1}^{n}\widehat{e}_i^2. \tag{4.31}$$

By the above calculation $\mathbb{E}\left[s^2 \mid \boldsymbol{X}\right] = \sigma^2$ and $\mathbb{E}\left[s^2\right] = \sigma^2$. Hence the estimator s^2 is unbiased for σ^2. Consequently, s^2 is known as the **bias-corrected estimator** for σ^2 and in empirical practice s^2 is the most widely used estimator for σ^2.

Interestingly, this is not the only method to construct an unbiased estimator for σ^2. An estimator constructed with the standardized residuals $\overline{e}_i$ from (4.28) is

$$\overline{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \overline{e}_i^2 = \frac{1}{n}\sum_{i=1}^{n} (1-h_{ii})^{-1}\widehat{e}_i^2.$$

You can show (see Exercise 4.9) that

$$\mathbb{E}\left[\overline{\sigma}^2 \mid \boldsymbol{X}\right] = \sigma^2 \tag{4.32}$$

and thus $\overline{\sigma}^2$ is unbiased for σ^2 (in the homoskedastic linear regression model).

When k/n is small the estimators $\widehat{\sigma}^2$, s^2 and $\overline{\sigma}^2$ are likely to be similar to one another. However, if k/n is large then s^2 and $\overline{\sigma}^2$ are generally preferred to $\widehat{\sigma}^2$. Consequently it is best to use one of the bias-corrected variance estimators in applications.

4.12 Mean-Square Forecast Error

One use of an estimated regression is to predict out-of-sample. Consider an out-of-sample realization (Y_{n+1}, X_{n+1}) where X_{n+1} is observed but not Y_{n+1}. Given the coefficient estimator $\widehat{\beta}$ the standard point estimator of $\mathbb{E}[Y_{n+1} \mid X_{n+1}] = X_{n+1}'\beta$ is $\widetilde{Y}_{n+1} = X_{n+1}'\widehat{\beta}$. The forecast error is the difference between the actual value Y_{n+1} and the point forecast $\widetilde{Y}_{n+1}$. This is the forecast error $\widetilde{e}_{n+1} = Y_{n+1} - \widetilde{Y}_{n+1}$. The mean-squared forecast error (MSFE) is its expected squared value $\mathrm{MSFE}_n = \mathbb{E}\left[\widetilde{e}_{n+1}^2\right]$. In the linear regression model $\widetilde{e}_{n+1} = e_{n+1} - X_{n+1}'\left(\widehat{\beta}-\beta\right)$ so

$$\mathrm{MSFE}_n = \mathbb{E}\left[e_{n+1}^2\right] - 2\mathbb{E}\left[e_{n+1}X_{n+1}'\left(\widehat{\beta}-\beta\right)\right] + \mathbb{E}\left[X_{n+1}'\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'X_{n+1}\right]. \tag{4.33}$$

The first term in (4.33) is σ^2. The second term in (4.33) is zero because $e_{n+1}X_{n+1}'$ is independent of $\widehat{\beta}-\beta$ and both are mean zero. Using the properties of the trace operator the third term in (4.33) is

$$\begin{aligned}
&\mathrm{tr}\left(\mathbb{E}\left[X_{n+1}X_{n+1}'\right]\mathbb{E}\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'\right]\right) \\
&= \mathrm{tr}\left(\mathbb{E}\left[X_{n+1}X_{n+1}'\right]\mathbb{E}\left[\mathbb{E}\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)' \mid \boldsymbol{X}\right]\right]\right) \\
&= \mathrm{tr}\left(\mathbb{E}\left[X_{n+1}X_{n+1}'\right]\mathbb{E}\left[\boldsymbol{V}_{\widehat{\beta}}\right]\right) \\
&= \mathbb{E}\left[\mathrm{tr}\left(\left(X_{n+1}X_{n+1}'\right)\boldsymbol{V}_{\widehat{\beta}}\right)\right] \\
&= \mathbb{E}\left[X_{n+1}'\boldsymbol{V}_{\widehat{\beta}}X_{n+1}\right]
\end{aligned} \tag{4.34}$$

where we use the fact that X_{n+1} is independent of $\widehat{\beta}$, the definition $\boldsymbol{V}_{\widehat{\beta}} = \mathbb{E}\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)' \mid \boldsymbol{X}\right]$, and the fact that X_{n+1} is independent of $\boldsymbol{V}_{\widehat{\beta}}$. Thus

$$\mathrm{MSFE}_n = \sigma^2 + \mathbb{E}\left[X_{n+1}'\boldsymbol{V}_{\widehat{\beta}}X_{n+1}\right].$$

Under conditional homoskedasticity this simplifies to

$$\mathrm{MSFE}_n = \sigma^2\left(1 + \mathbb{E}\left[X_{n+1}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}X_{n+1}\right]\right).$$

A simple estimator for the MSFE is obtained by averaging the squared prediction errors (3.46)

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

where $\widetilde{e}_i = Y_i - X_i'\widehat{\beta}_{(-i)} = \widehat{e}_i(1-h_{ii})^{-1}$. Indeed, we can calculate that

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

By a similar calculation as in (4.34) we find

$$\mathbb{E}\left[\widetilde{\sigma}^2\right] = \sigma^2 + \mathbb{E}\left[X_i' \boldsymbol{V}_{\widehat{\beta}_{(-i)}} X_i\right] = \mathrm{MSFE}_{n-1}.$$

This is the MSFE based on a sample of size $n-1$ rather than size n. The difference arises because the in-sample prediction errors $\widetilde{e}_i$ for $i \le n$ are calculated using an effective sample size of $n-1$, while the out-of sample prediction error $\widetilde{e}_{n+1}$ is calculated from a sample with the full n observations. Unless n is very small we should expect MSFE_{n-1} (the MSFE based on $n-1$ observations) to be close to MSFE_n (the MSFE based on n observations). Thus $\widetilde{\sigma}^2$ is a reasonable estimator for MSFE_n.

Theorem 4.6 MSFE
In the linear regression model (Assumption 4.2) and i.i.d. sampling (Assumption 4.1)

$$\mathrm{MSFE}_n = \mathbb{E}\left[\widetilde{e}_{n+1}^2\right] = \sigma^2 + \mathbb{E}\left[X_{n+1}' \boldsymbol{V}_{\widehat{\beta}} X_{n+1}\right]$$

where $\boldsymbol{V}_{\widehat{\beta}} = \mathrm{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right]$. Furthermore, $\widetilde{\sigma}^2$ defined in (3.46) is an unbiased estimator of MSFE_{n-1}, because $\mathbb{E}\left[\widetilde{\sigma}^2\right] = \mathrm{MSFE}_{n-1}$.

4.13 Covariance Matrix Estimation Under Homoskedasticity

For inference we need an estimator of the covariance matrix $\boldsymbol{V}_{\widehat{\beta}}$ of the least squares estimator. In this section we consider the homoskedastic regression model (Assumption 4.3).

Under homoskedasticity the covariance matrix takes the simple form

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

which is known up to the scale σ^2. In Section 4.11 we discussed three estimators of σ^2. The most commonly used choice is s^2 leading to the classic covariance matrix estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}}^0 = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} s^2. \tag{4.35}$$

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

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

This was the dominant covariance matrix estimator in applied econometrics for many years and is still the default method in most regression packages. For example, Stata uses the covariance matrix estimator (4.35) by default in linear regression unless an alternative is specified.

If the estimator (4.35) is used but the regression error is heteroskedastic it is possible for $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0}$ to be quite biased for the correct covariance matrix $\boldsymbol{V}_{\widehat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$. For example, suppose $k = 1$ and $\sigma_i^2 = X_i^2$ with $\mathbb{E}[X] = 0$. The ratio of the true variance of the least squares estimator to the expectation of the variance estimator is

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

(Notice that we use the fact that $\sigma_i^2 = X_i^2$ implies $\sigma^2 = \mathbb{E}\left[\sigma_i^2\right] = \mathbb{E}\left[X^2\right]$.) The constant κ is the standardized fourth moment (or kurtosis) of the regressor X and can be any number greater than one. For example, if $X \sim \mathrm{N}\left(0, \sigma^2\right)$ then $\kappa = 3$, so the true variance $\boldsymbol{V}_{\widehat{\beta}}$ is three times larger than the expected homoskedastic estimator $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0}$. But κ can be much larger. Take, for example, the variable *wage* in the CPS data set. It satisfies $\kappa = 30$ so that if the conditional variance equals $\sigma_i^2 = X_i^2$ then the true variance $\boldsymbol{V}_{\widehat{\beta}}$ is 30 times larger than the expected homoskedastic estimator $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0}$. While this is an extreme case the point is that the classic covariance matrix estimator (4.35) may be quite biased when the homoskedasticity assumption fails.

4.14 Covariance Matrix Estimation Under Heteroskedasticity

In the previous section we showed that that the classic covariance matrix estimator can be highly biased if homoskedasticity fails. In this section we show how to construct covariance matrix estimators which do not require homoskedasticity.

Recall that the general form for the covariance matrix is

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

with $\boldsymbol{D}$ defined in (4.8). This depends on the unknown matrix $\boldsymbol{D}$ which we can write as

$$\boldsymbol{D} = \operatorname{diag}\left(\sigma_1^2, ..., \sigma_n^2\right) = \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right] = \mathbb{E}\left[\widetilde{\boldsymbol{D}} \mid \boldsymbol{X}\right]$$

where $\widetilde{\boldsymbol{D}} = \operatorname{diag}\left(e_1^2, ..., e_n^2\right)$. Thus $\widetilde{\boldsymbol{D}}$ is a conditionally unbiased estimator for $\boldsymbol{D}$. If the squared errors e_i^2 were observable, we could construct an unbiased estimator for $\boldsymbol{V}_{\widehat{\beta}}$ as

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

Indeed,

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

verifying that $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{ideal}}$ is unbiased for $\boldsymbol{V}_{\widehat{\beta}}$.

Since the errors e_i^2 are unobserved $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{ideal}}$ is not a feasible estimator. However, we can replace e_i^2 with the squared residuals $\widehat{e}_i^2$. Making this substitution we obtain the estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC0}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.36}$$

The label "HC" refers to "heteroskedasticity-consistent". The label "HC0" refers to this being the baseline heteroskedasticity-consistent covariance matrix estimator.

We know, however, that $\widehat{e}_i^2$ is biased towards zero (recall equation (4.27)). To estimate the variance σ^2 the unbiased estimator s^2 scales the moment estimator $\widehat{\sigma}^2$ by $n/(n-k)$. Making the same adjustment we obtain the estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC1}} = \left(\frac{n}{n-k}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.37}$$

While the scaling by $n/(n-k)$ is *ad hoc*, HC1 is often recommended over the unscaled HC0 estimator.

Alternatively, we could use the standardized residuals $\overline{e}_i$ or the prediction errors $\widetilde{e}_i$, yielding the "HC2" and "HC3" estimators

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC2}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \overline{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} (1-h_{ii})^{-1} X_i X_i' \widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\end{aligned} \tag{4.38}$$

and

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC3}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widetilde{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} (1-h_{ii})^{-2} X_i X_i' \widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}.\end{aligned} \tag{4.39}$$

The four estimators HC0, HC1, HC2, and HC3 are collectively called **robust**, **heteroskedasticity-consistent**, or **heteroskedasticity-robust** covariance matrix estimators. The HC0 estimator was first developed by Eicker (1963) and introduced to econometrics by White (1980) and is sometimes called the **Eicker-White** or **White** covariance matrix estimator. The degree-of-freedom adjustment in HC1 was recommended by Hinkley (1977) and is the default robust covariance matrix estimator implemented in Stata. It is implement by the ", `r`" option. In current applied econometric practice this is the most popular covariance matrix estimator. The HC2 estimator was introduced by Horn, Horn and Duncan (1975) and is implemented using the `vce(hc2)` option in Stata. The HC3 estimator was derived by MacKinnon and White (1985) from the jackknife principle (see Section 10.3), and by Andrews (1991a) based on the principle of leave-one-out cross-validation, and is implemented using the `vce(hc3)` option in Stata.

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

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC0}} < \widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC2}} < \widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC3}}. \tag{4.40}$$

(See Exercise 4.10.) The inequality $\boldsymbol{A} < \boldsymbol{B}$ when applied to matrices means that the matrix $\boldsymbol{B} - \boldsymbol{A}$ is positive definite.

In general, the bias of the covariance matrix estimators is complicated but simplify under the assumption of homoskedasticity (4.3). For example, using (4.27),

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}} \mid \boldsymbol{X}\right] &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \mathbb{E}\left[\widehat{e}_i^2 \mid \boldsymbol{X}\right]\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \left(1-h_{ii}\right)\sigma^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' h_{ii}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 \\
&< \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2 = \boldsymbol{V}_{\widehat{\beta}}.
\end{aligned}$$

This calculation shows that $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$ is biased towards zero.

By a similar calculation (again under homoskedasticity) we can calculate that the HC2 estimator is unbiased

$$\mathbb{E}\left[\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC2}} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^2. \tag{4.41}$$

(See Exercise 4.11.)

It might seem rather odd to compare the bias of heteroskedasticity-robust estimators under the assumption of homoskedasticity but it does give us a baseline for comparison.

Another interesting calculation shows that in general (that is, without assuming homoskedasticity) the HC3 estimator is biased away from zero. Indeed, using the definition of the prediction errors (3.44)

$$\widetilde{e}_i = Y_i - X_i'\widehat{\beta}_{(-i)} = e_i - X_i'\left(\widehat{\beta}_{(-i)} - \beta\right)$$

so

$$\widetilde{e}_i^2 = e_i^2 - 2X_i'\left(\widehat{\beta}_{(-i)} - \beta\right)e_i + \left(X_i'\left(\widehat{\beta}_{(-i)} - \beta\right)\right)^2.$$

Note that e_i and $\widehat{\beta}_{(-i)}$ are functions of non-overlapping observations and are thus independent. Hence $\mathbb{E}\left[\left(\widehat{\beta}_{(-i)} - \beta\right)e_i \mid \boldsymbol{X}\right] = 0$ and

$$\begin{aligned}
\mathbb{E}\left[\widetilde{e}_i^2 \mid \boldsymbol{X}\right] &= \mathbb{E}\left[e_i^2 \mid \boldsymbol{X}\right] - 2X_i'\mathbb{E}\left[\left(\widehat{\beta}_{(-i)} - \beta\right)e_i \mid \boldsymbol{X}\right] + \mathbb{E}\left[\left(X_i'\left(\widehat{\beta}_{(-i)} - \beta\right)\right)^2 \mid \boldsymbol{X}\right] \\
&= \sigma_i^2 + \mathbb{E}\left[\left(X_i'\left(\widehat{\beta}_{(-i)} - \beta\right)\right)^2 \mid \boldsymbol{X}\right] \\
&\geq \sigma_i^2.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}\left[\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC3}} \mid \boldsymbol{X}\right] &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \mathbb{E}\left[\widetilde{e}_i^2 \mid \boldsymbol{X}\right]\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&\geq \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \sigma_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \boldsymbol{V}_{\widehat{\beta}}.
\end{aligned}$$

This means that the HC3 estimator is conservative in the sense that it is weakly larger (in expectation) than the correct variance for any realization of $\boldsymbol{X}$.

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

Of the four robust estimators HC1 is the most commonly used as it is the default robust covariance matrix option in Stata. However, HC2 and HC3 are preferred. HC2 is unbiased (under homoskedasticity) and HC3 is conservative for any $\boldsymbol{X}$. In most applications HC1, HC2, and HC3 will be similar so this choice will not matter. The context where the estimators can differ substantially is when the sample has a large leverage value h_{ii} for at least one observation. You can see this by comparing the formulas (4.37), (4.38), and (4.39) and noting that the only difference is the scaling by the leverage values h_{ii}. If there is an observation with h_{ii} close to one, then $(1-h_{ii})^{-1}$ and $(1-h_{ii})^{-2}$ will be large, giving this observation much greater weight in the covariance matrix formula.

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 is one of the most cited papers 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. He also pioneered the introduction of neural network methods into econometrics. In addition to being a highly prolific and influential scholar, he also co-founded the economic consulting firm Bates White.

4.15 Standard Errors

A variance estimator such as $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ is an estimator of the variance of the distribution of $\widehat{\beta}$. A more easily interpretable measure of spread is its square root – the standard deviation. This is so important when discussing the distribution of parameter estimators we have a special name for estimates of their standard deviation.

Definition 4.2 A **standard error** $s(\widehat{\beta})$ for a real-valued estimator $\widehat{\beta}$ is an estimator of the standard deviation of the distribution of $\widehat{\beta}$.

When β is a vector with estimator $\widehat{\beta}$ and covariance matrix estimator $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$, standard errors for individual elements are the square roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$. That is,

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

When the classical covariance matrix estimator (4.35) is used the standard error takes the simple form

$$s(\widehat{\beta}_j) = s\sqrt{\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}}. \tag{4.42}$$

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 (3.12) of Section 3.7. We calculate that $s^2 = 0.160$. Therefore the homoskedastic covariance matrix estimate is

$$\widehat{\boldsymbol{V}}^0_{\widehat{\beta}} = \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} 0.160 = \begin{pmatrix} 0.002 & -0.031 \\ -0.031 & 0.499 \end{pmatrix}.$$

We also calculate that

$$\sum_{i=1}^{n} (1 - h_{ii})^{-1} X_i X_i' \widehat{e}_i^2 = \begin{pmatrix} 763.26 & 48.513 \\ 48.513 & 3.1078 \end{pmatrix}.$$

Therefore the HC2 covariance matrix estimate is

$$\begin{aligned} \widehat{\boldsymbol{V}}^{\mathrm{HC2}}_{\widehat{\beta}} &= \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \begin{pmatrix} 763.26 & 48.513 \\ 48.513 & 3.1078 \end{pmatrix} \begin{pmatrix} 5010 & 314 \\ 314 & 20 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} 0.001 & -0.015 \\ -0.015 & 0.243 \end{pmatrix}. \end{aligned} \tag{4.43}$$

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

$$\begin{aligned} \widehat{\log(wage)} = \underset{(0.031)}{0.155} \quad education + \underset{(0.493)}{0.698} \quad . \end{aligned} \tag{4.44}$$

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

Table 4.1: Standard Errors

	Education	Intercept
Homoskedastic (4.35)	0.045	0.707
HC0 (4.36)	0.029	0.461
HC1 (4.37)	0.030	0.486
HC2 (4.38)	0.031	0.493
HC3 (4.39)	0.033	0.527

The homoskedastic standard errors are noticeably different (larger in this case) than the others. The robust standard errors are reasonably close to one another though the HC3 standard errors are larger than the others.

4.16 Estimation with Sparse Dummy Variables

The heteroskedasticity-robust covariance matrix estimators can be quite imprecise in some contexts. One is in the presence of **sparse dummy variables** – when a dummy variable only takes the value 1 or 0 for very few observations. In these contexts one component of the covariance matrix is estimated on just those few observations and will be imprecise. This is effectively hidden from the user.

To see the problem, let D be a dummy variable (takes on the values 1 and 0) and consider the dummy variable regression

$$Y = \beta_1 D + \beta_2 + e. \tag{4.45}$$

The number of observations for which $D_i = 1$ is $n_1 = \sum_{i=1}^n D_i$. The number of observations for which $D_i = 0$ is $n_2 = n - n_1$. We say the design is **sparse** if n_1 or n_2 is small.

To simplify our analysis, we take the extreme case $n_1 = 1$. The ideas extend to the case of $n_1 > 1$ but small, though with less dramatic effects.

In the regression model (4.45) we can calculate that the true covariance matrix of the least squares estimator for the coefficients under the simplifying assumption of conditional homoskedasticity is

$$\boldsymbol{V}_{\widehat{\beta}} = \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} = \sigma^2 \begin{pmatrix} 1 & 1 \\ 1 & n \end{pmatrix}^{-1} = \frac{\sigma^2}{n-1} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix}.$$

In particular, the variance of the estimator for the coefficient on the dummy variable is

$$V_{\widehat{\beta}_1} = \sigma^2 \frac{n}{n-1}.$$

Essentially, the coefficient β_1 is estimated from a single observation so its variance is roughly unaffected by sample size. An important message is that certain coefficient estimators in the presence of sparse dummy variables will be imprecise, regardless of the sample size. A large sample alone is not sufficient to ensure precise estimation.

Now let's examine the standard HC1 covariance matrix estimator (4.37). The regression has perfect fit for the observation for which $D_i = 1$ so the corresponding residual is $\widehat{e}_i = 0$. It follows that $D_i \widehat{e}_i = 0$ for all i (either $D_i = 0$ or $\widehat{e}_i = 0$). Hence

$$\sum_{i=1}^n X_i X_i' \widehat{e}_i^2 = \begin{pmatrix} 0 & 0 \\ 0 & \sum_{i=1}^n \widehat{e}_i^2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & (n-2)s^2 \end{pmatrix}$$

where $s^2 = (n-2)^{-1} \sum_{i=1}^n \widehat{e}_i^2$ is the bias-corrected estimator of σ^2. Together we find that

$$\begin{aligned} \widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC1}} &= \left(\frac{n}{n-2}\right) \frac{1}{(n-1)^2} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & (n-2)s^2 \end{pmatrix} \begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix} \\ &= s^2 \frac{n}{(n-1)^2} \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}. \end{aligned}$$

In particular, the estimator for $V_{\widehat{\beta}_1}$ is

$$\widehat{V}_{\widehat{\beta}_1}^{\mathrm{HC1}} = s^2 \frac{n}{(n-1)^2}.$$

It has expectation

$$\mathbb{E}\left[\widehat{V}_{\widehat{\beta}_1}^{\mathrm{HC1}}\right] = \sigma^2 \frac{n}{(n-1)^2} = \frac{V_{\widehat{\beta}_1}}{n-1} << V_{\widehat{\beta}_1}.$$

The variance estimator $\widehat{V}_{\widehat{\beta}_1}^{\mathrm{HC1}}$ is extremely biased for $V_{\widehat{\beta}_1}$. It is too small by a multiple of n! The reported variance – and standard error – is misleadingly small. The variance estimate erroneously mis-states the precision of $\widehat{\beta}_1$.

The fact that $\widehat{V}_{\widehat{\beta}_1}^{\mathrm{HC1}}$ is biased is unlikely to be noticed by an applied researcher. Nothing in the reported output will alert a researcher to the problem.

Another way to see the issue is to consider the estimator $\widehat{\theta} = \widehat{\beta}_1 + \widehat{\beta}_2$ for the sum of the coefficients $\theta = \beta_1 + \beta_2$. This estimator has true variance σ^2. The variance estimator, however is $\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{\text{HC1}} = 0$! (It equals the sum of the four elements in $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC1}}$). Clearly, the estimator "0" is biased for the true value σ^2.

Another insight is to examine the leverage values. The (single) observation with $D_i = 1$ has

$$h_{ii} = \frac{1}{n-1}\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} n & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 1.$$

This is an extreme leverage value.

A possible solution is to replace the biased covariance matrix estimator $\widehat{V}_{\widehat{\beta}_1}^{\text{HC1}}$ with the unbiased estimator $\widehat{V}_{\widehat{\beta}_1}^{\text{HC2}}$ (unbiased under homoskedasticity) or the conservative estimator $\widehat{V}_{\widehat{\beta}_1}^{\text{HC3}}$. Neither approach can be done in the extreme sparse case $n_1 = 1$ (for $\widehat{V}_{\widehat{\beta}_1}^{\text{HC2}}$ and $\widehat{V}_{\widehat{\beta}_1}^{\text{HC3}}$ cannot be calculated if $h_{ii} = 1$ for any observation) but applies otherwise. When $h_{ii} = 1$ for an observation then $\widehat{V}_{\widehat{\beta}_1}^{\text{HC2}}$ and $\widehat{V}_{\widehat{\beta}_1}^{\text{HC3}}$ cannot be calculated. In this case unbiased covariance matrix estimation appears to be impossible.

It is unclear if there is a best practice to avoid this situation. Once possibility is to calculate the maximum leverage value. If it is very large calculate the standard errors using several methods to see if variation occurs.

4.17 Computation

We illustrate methods to compute standard errors for equation (3.13) extending the code of Section 3.25.

Stata do File (continued)

```
*     Homoskedastic formula (4.35):
reg wage education experience exp2 if (mnwf == 1)
*     HC1 formula (4.37):
reg wage education experience exp2 if (mnwf == 1), r
*     HC2 formula (4.38):
reg wage education experience exp2 if (mnwf == 1), vce(hc2)
*     HC3 formula (4.39):
reg wage education experience exp2 if (mnwf == 1), vce(hc3)
```

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/sqrt(1-leverage))%*%matrix(1,1,k))
u3 <- x*((e/(1-leverage))%*%matrix(1,1,k))
xx <- solve(t(x)%*%x)
v0 <- xx*sig2
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))            # HC0
s1a <- sqrt(diag(v1a))          # HC1
s2 <- sqrt(diag(v2))            # HC2
s3 <- sqrt(diag(v3))            # HC3
```

MATLAB Program File (continued)

```
[n,k]=size(x);
a=n/(n-k);
sig2=(e'*e)/(n-k);
u1=x.*e;u2=x.*(e./sqrt(1-leverage));
u3=x.*(e./(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));              # HC0 formula
s1a=sqrt(diag(v1a));            # HC1 formula
s2=sqrt(diag(v2));              # HC2 formula
s3=sqrt(diag(v3));              # HC3 formula
```

4.18 Measures of Fit

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

$$R^2 = 1 - \frac{\sum_{i=1}^n \widehat{e}_i^2}{\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}_Y^2}.$$

where $\widehat{\sigma}_Y^2 = n^{-1}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2$. R^2 is an estimator of the population parameter

$$\rho^2 = \frac{\text{var}\left[X'\beta\right]}{\text{var}\left[Y\right]} = 1 - \frac{\sigma^2}{\sigma_Y^2}.$$

However, $\widehat{\sigma}^2$ and $\widehat{\sigma}_Y^2$ are biased. Theil (1961) proposed replacing these by the unbiased versions s^2 and $\widetilde{\sigma}_Y^2 = (n-1)^{-1}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2$ yielding what is known as **R-bar-squared** or **adjusted R-squared**:

$$\overline{R}^2 = 1 - \frac{s^2}{\widetilde{\sigma}_Y^2} = 1 - \frac{(n-1)^{-1}\sum_{i=1}^n \widehat{e}_i^2}{(n-k)^{-1}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2}.$$

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

$$\widetilde{R}^2 = 1 - \frac{\sum_{i=1}^n \widetilde{e}_i^2}{\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\widetilde{\sigma}^2}{\widehat{\sigma}_Y^2}$$

where $\widetilde{e}_i$ are the prediction errors (3.44) and $\widetilde{\sigma}^2$ is the MSPE from (3.46). As described in Section (4.12) $\widetilde{\sigma}^2$ is a good estimator of the out-of-sample mean-squared forecast error so $\widetilde{R}^2$ is a good estimator of the percentage of the forecast variance which is explained by the regression forecast. In this sense $\widetilde{R}^2$ is a good measure of fit.

One problem with R^2 which is partially corrected by $\overline{R}^2$ and fully corrected by $\widetilde{R}^2$ is that R^2 necessarily increases when regressors are added to a regression model. This occurs because R^2 is a negative function of the sum of squared residuals which cannot increase when a regressor is added. In contrast, $\overline{R}^2$ and $\widetilde{R}^2$ are non-monotonic in the number of regressors. $\widetilde{R}^2$ can even be negative, which occurs when an estimated model predicts worse than a constant-only model.

In the statistical literature the MSPE $\widetilde{\sigma}^2$ is known as the **leave-one-out cross validation** criterion and is popular for model comparison and selection, especially in high-dimensional and nonparametric contexts. It is equivalent to use $\widetilde{R}^2$ or $\widetilde{\sigma}^2$ to compare and select models. Models with high $\widetilde{R}^2$ (or low $\widetilde{\sigma}^2$) are better models in terms of expected out of sample squared error. In contrast, R^2 cannot be used for model selection as it necessarily increases when regressors are added to a regression model. $\overline{R}^2$ is also an inappropriate choice for model selection (it tends to select models with too many parameters) though a justification of this assertion requires a study of the theory of model selection. Unfortunately, $\overline{R}^2$ is routinely used by some economists, possibly as a hold-over from previous generations.

In summary, it is recommended to omit R^2 and $\overline{R}^2$. If a measure of fit is desired, report $\widetilde{R}^2$ or $\widetilde{\sigma}^2$.

Henri Theil

Henri Theil (1924-2000) of the Netherlands invented $\overline{R}^2$ and two-stage least squares, both of which are routinely seen in applied econometrics. He also wrote an early influential advanced textbook on econometrics (Theil, 1971).

4.19 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 4.2. For standard errors we use the unbiased HC2 formula.

Table 4.2 displays the parameter estimates in a standard tabular format. 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 4.2: OLS Estimates of Linear Equation for log(wage)

	$\widehat{\beta}$	$s(\widehat{\beta})$
Education	0.117	0.001
Experience	0.033	0.001
Experience^2/100	-0.056	0.002
Female	-0.098	0.011
Female Union Member	0.023	0.020
Male Union Member	0.095	0.020
Married Female	0.016	0.010
Married Male	0.211	0.010
Formerly Married Female	-0.006	0.012
Formerly Married Male	0.083	0.015
Hispanic	-0.108	0.008
Black	-0.096	0.008
American Indian	-0.137	0.027
Asian	-0.038	0.013
Mixed Race	-0.041	0.021
Intercept	0.909	0.021
$\widehat{\sigma}$	0.565	
Sample Size	46,943	

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 4.2 confirm our earlier findings that the return to a year of education is approximately 12%, the return to experience is concave, 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

[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.

are wage premiums for men who are members of a labor union (about 10%), married (about 21%) or formerly married (about 8%), but no similar premiums are apparent for women.

4.20 Multicollinearity

As discussed in Section 3.24, if $\boldsymbol{X}'\boldsymbol{X}$ is singular then $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and $\widehat{\beta}$ are not defined. This situation is called **strict multicollinearity** as the columns of $\boldsymbol{X}$ are linearly dependent, i.e., there is some $\alpha \neq 0$ such that $\boldsymbol{X}\alpha = 0$. Most commonly this arises when sets of regressors are included which are identically related. In Section 3.24 we discussed possible causes of strict multicollinearity and discussed the related problem of ill-conditioning which can cause numerical inaccuracies in severe cases.

A related common situation is **near multicollinearity** which is often called "multicollinearity" for brevity. This is the situation when the regressors are highly correlated. An implication of near multicollinearity is that individual coefficient estimates will be imprecise. This is not necessarily a problem for econometric analysis if the reported standard errors are accurate. However, robust standard errors can be sensitive to large leverage values which can occur under near multicollinearity. This leads to the undesirable situation where the coefficient estimates are imprecise yet the standard errors are misleadingly small.

We can see the impact of near multicollinearity on precision in a simple homoskedastic linear regression model with two regressors

$$Y = X_1\beta_1 + X_2\beta_2 + e,$$

and

$$\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

In this case

$$\operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \frac{\sigma^2}{n}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n\left(1-\rho^2\right)}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

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.

Many early-generation textbooks overemphasized multicollinearity. An amusing parody of these texts is *Micronumerosity*, Chapter 23.3 of Goldberger's *A Course in Econometrics* (1991). Among the witty remarks of his chapter are the following.

> 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.)
>
> 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.

Arthur S. Goldberger, *A Course in Econometrics* (1991), pp. 249.

To understand Goldberger's basic point you should notice that 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. He was pointing out that the only statistical implication of multicollinearity in the homoskedastic model is a lack of precision. Small sample sizes have the exact same implication.

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 Ph.D. thesis developed a pioneering macroeconometric forecasting model (the Klein-Goldberger model). Most of his remaining career focused on microeconometric issues. He was the leading pioneer of what has been called the Wisconsin Tradition of empirical work – a combination of formal econometric theory with a careful critical analysis of empirical work. Goldberger wrote a series of highly regarded and influential graduate econometric textbooks, including *Econometric Theory* (1964), *Topics in Regression Analysis* (1968), and *A Course in Econometrics* (1991).

4.21 Clustered Sampling

In Section 4.2 we briefly mentioned clustered sampling as an alternative to the assumption of random sampling. We now introduce the framework in more detail and extend the primary results of this chapter to encompass clustered dependence.

It might be easiest to understand the idea of clusters by considering a concrete example. Duflo, Dupas, and Kremer (2011) investigate the impact of tracking (assigning students based on initial test score) on educational attainment in a randomized experiment. An extract of their data set is available on the textbook webpage in the file `DDK2011`.

In 2005, 140 primary schools in Kenya received funding to hire an extra first grade teacher to reduce class sizes. In half of the schools (selected randomly) students were assigned to classrooms based on an initial test score ("tracking"); in the remaining schools the students were randomly assigned to classrooms. For their analysis the authors restricted attention to the 121 schools which initially had a single first-grade class.

The key regression[4] in the paper is

$$TestScore_{ig} = -0.071 + 0.138 Tracking_g + e_{ig} \tag{4.46}$$

where $TestScore_{ig}$ is the standardized test score (normalized to have mean 0 and variance 1) of student i in school g, and $Tracking_g$ is a dummy equal to 1 if school g was tracking. The OLS estimates indicate that schools which tracked the students had an overall increase in test scores by about 0.14 standard deviations, which is meaningful. More general versions of this regression are estimated, many of which take the form

$$TestScore_{ig} = \alpha + \gamma Tracking_g + X_{ig}'\beta + e_{ig} \tag{4.47}$$

where X_{ig} is a set of controls specific to the student (including age, gender, and initial test score).

[4] Table 2, column (1). Duflo, Dupas and Kremer (2011) report a coefficient estimate of 0.139, perhaps due to a slightly different calculation to standardize the test score.

A difficulty with applying the classical regression framework is that student achievement is likely correlated within a given school. Student achievement may be affected by local demographics, individual teachers, and classmates, all of which imply dependence. These concerns, however, do not suggest that achievement will be correlated across schools, so it seems reasonable to model achievement across schools as mutually independent. We call such dependence **clustered**.

In clustering contexts it is convenient to double index the observations as (Y_{ig}, X_{ig}) where $g = 1, ..., G$ indexes the cluster and $i = 1, ..., n_g$ indexes the individual within the g^{th} cluster. The number of observations per cluster n_g may vary across clusters. The number of clusters is G. The total number of observations is $n = \sum_{g=1}^{G} n_g$. In the Kenyan schooling example the number of clusters (schools) in the estimation sample is $G = 121$, the number of students per school varies from 19 to 62, and the total number of observations is $n = 5795$.

While it is typical to write the observations using the double index notation (Y_{ig}, X_{ig}) it is also useful to use cluster-level notation. Let $\boldsymbol{Y}_g = (Y_{1g}, ..., Y_{n_g g})'$ and $\boldsymbol{X}_g = (X_{1g}, ..., X_{n_g g})'$ denote the $n_g \times 1$ vector of dependent variables and $n_g \times k$ matrix of regressors for the g^{th} cluster. A linear regression model can be written by individual as

$$Y_{ig} = X_{ig}'\beta + e_{ig}$$

and using cluster notation as

$$\boldsymbol{Y}_g = \boldsymbol{X}_g\beta + \boldsymbol{e}_g \tag{4.48}$$

where $\boldsymbol{e}_g = (e_{1g}, ..., e_{n_g g})'$ is a $n_g \times 1$ error vector. We can also stack the observations into full sample matrices and write the model as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{e}.$$

Using this notation we can write the sums over the observations using the double sum $\sum_{g=1}^{G}\sum_{i=1}^{n_g}$. This is the sum across clusters of the sum across observations within each cluster. The OLS estimator can be written as

$$\begin{aligned}\widehat{\beta} &= \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}X_{ig}'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}Y_{ig}\right) \\ &= \left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{X}_g\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\boldsymbol{Y}_g\right) \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right).\end{aligned} \tag{4.49}$$

The residuals are $\widehat{e}_{ig} = Y_{ig} - X_{ig}'\widehat{\beta}$ in individual level notation and $\widehat{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g\widehat{\beta}$ in cluster level notation.

The standard clustering assumption is that the clusters are known to the researcher and that the observations are independent across clusters.

Assumption 4.4 The clusters $(\boldsymbol{Y}_g, \boldsymbol{X}_g)$ are mutually independent across clusters g.

In our example clusters are schools. In other common applications cluster dependence has been assumed within individual classrooms, families, villages, regions, and within larger units such as industries and states. This choice is up to the researcher though the justification will depend on the context, the nature of the data, and will reflect information and assumptions on the dependence structure across observations.

The model is a linear regression under the assumption

$$\mathbb{E}\left[\boldsymbol{e}_g \mid \boldsymbol{X}_g\right] = 0. \tag{4.50}$$

This is the same as assuming that the individual errors are conditionally mean zero

$$\mathbb{E}\left[e_{ig} \mid \boldsymbol{X}_g\right] = 0$$

or that the conditional expectation of $\boldsymbol{Y}_g$ given $\boldsymbol{X}_g$ is linear. As in the independent case equation (4.50) means that the linear regression model is correctly specified. In the clustered regression model this requires that all interaction effects within clusters have been accounted for in the specification of the individual regressors X_{ig}.

In the regression (4.46) the conditional expectation is necessarily linear and satisfies (4.50) since the regressor $Tracking_g$ is a dummy variable at the cluster level. In the regression (4.47) with individual controls, (4.50) requires that the achievement of any student is unaffected by the individual controls (e.g. age, gender, and initial test score) of other students within the same school.

Given (4.50) we can calculate the expectation of the OLS estimator. Substituting (4.48) into (4.49) we find

$$\widehat{\beta} - \beta = \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right).$$

The mean of $\widehat{\beta} - \beta$ conditioning on all the regressors is

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

The first equality holds by linearity, the second by Assumption 4.4, and the third by (4.50).

This shows that OLS is unbiased under clustering if the conditional expectation is linear.

Theorem 4.7 In the clustered linear regression model (Assumption 4.4 and (4.50)) $\mathbb{E}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \beta$.

Now consider the covariance matrix of $\widehat{\beta}$. Let $\Sigma_g = \mathbb{E}\left[\boldsymbol{e}_g \boldsymbol{e}_g' \mid \boldsymbol{X}_g\right]$ denote the $n_g \times n_g$ conditional covariance matrix of the errors within the g^{th} cluster. Since the observations are independent across clusters,

$$\begin{aligned}
\operatorname{var}\left[\left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{e}_g\right) \middle| \boldsymbol{X}\right] &= \sum_{g=1}^{G} \operatorname{var}\left[\boldsymbol{X}_g' \boldsymbol{e}_g \mid \boldsymbol{X}_g\right] \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' \mathbb{E}\left[\boldsymbol{e}_g \boldsymbol{e}_g' \mid \boldsymbol{X}_g\right] \boldsymbol{X}_g \\
&= \sum_{g=1}^{G} \boldsymbol{X}_g' \Sigma_g \boldsymbol{X}_g \\
&\stackrel{\text{def}}{=} \Omega_n.
\end{aligned} \tag{4.51}$$

It follows that

$$\boldsymbol{V}_{\widehat{\beta}} = \operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \Omega_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.52}$$

This differs from the formula in the independent case due to the correlation between observations within clusters. The magnitude of the difference depends on the degree of correlation between observations within clusters and the number of observations within clusters. To see this, suppose that all clusters have the same number of observations $n_g = N$, $\mathbb{E}\left[e_{ig}^2 \mid \boldsymbol{X}_g\right] = \sigma^2$, $\mathbb{E}\left[e_{ig}e_{\ell g} \mid \boldsymbol{X}_g\right] = \sigma^2\rho$ for $i \neq \ell$, and the regressors X_{ig} do not vary within a cluster. In this case the exact variance of the OLS estimator equals[5] (after some calculations)

$$\boldsymbol{V}_{\widehat{\beta}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \sigma^2 \left(1 + \rho\left(N - 1\right)\right). \tag{4.53}$$

If $\rho > 0$ the exact variance is appropriately a multiple ρN of the conventional formula. In the Kenyan school example the average cluster size is 48. If $\rho = 0.25$ this means the exact variance exceeds the conventional formula by a factor of about twelve. In this case the correct standard errors (the square root of the variance) are a multiple of about three times the conventional formula. This is a substantial difference and should not be neglected.

Arellano (1987) proposed a cluster-robust covariance matrix estimator which is an extension of the White estimator. Recall that the insight of the White covariance estimator is that the squared error e_i^2 is unbiased for $\mathbb{E}\left[e_i^2 \mid X_i\right] = \sigma_i^2$. Similarly, with cluster dependence the matrix $\boldsymbol{e}_g\boldsymbol{e}_g'$ is unbiased for $\mathbb{E}\left[\boldsymbol{e}_g\boldsymbol{e}_g' \mid \boldsymbol{X}_g\right] = \Sigma_g$. This means that an unbiased estimator for (4.51) is $\widetilde{\Omega}_n = \sum_{g=1}^{G} \boldsymbol{X}_g'\boldsymbol{e}_g\boldsymbol{e}_g'\boldsymbol{X}_g$. This is not feasible, but we can replace the unknown errors by the OLS residuals to obtain Arellano's estimator:

$$\begin{aligned}
\widehat{\Omega}_n &= \sum_{g=1}^{G} \boldsymbol{X}_g'\widehat{\boldsymbol{e}}_g\widehat{\boldsymbol{e}}_g'\boldsymbol{X}_g \\
&= \sum_{g=1}^{G} \sum_{i=1}^{n_g} \sum_{\ell=1}^{n_g} X_{ig}X_{\ell g}'\widehat{e}_{ig}\widehat{e}_{\ell g} \\
&= \sum_{g=1}^{G} \left(\sum_{i=1}^{n_g} X_{ig}\widehat{e}_{ig}\right)\left(\sum_{\ell=1}^{n_g} X_{\ell g}\widehat{e}_{\ell g}\right)'.
\end{aligned} \tag{4.54}$$

The three expressions in (4.54) give three equivalent formulae which could be used to calculate $\widehat{\Omega}_n$. The final expression writes $\widehat{\Omega}_n$ in terms of the cluster sums $\sum_{\ell=1}^{n_g} X_{\ell g}\widehat{e}_{\ell g}$ which is the basis for our example R and MATLAB codes shown below.

Given the expressions (4.51)-(4.52) a natural cluster covariance matrix estimator takes the form

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}} = a_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \widehat{\Omega}_n \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tag{4.55}$$

where a_n is a possible finite-sample adjustment. The Stata `cluster` command uses

$$a_n = \left(\frac{n-1}{n-k}\right)\left(\frac{G}{G-1}\right). \tag{4.56}$$

The factor $G/(G-1)$ was derived by Chris Hansen (2007) in the context of equal-sized clusters to improve performance when the number of clusters G is small. The factor $(n-1)/(n-k)$ is an *ad hoc* generalization which nests the adjustment used in (4.37) since $G = n$ implies the simplification $a_n = n/(n-k)$.

Alternative cluster-robust covariance matrix estimators can be constructed using cluster-level prediction errors such as $\widetilde{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g\widehat{\beta}_{(-g)}$ where $\widehat{\beta}_{(-g)}$ is the least squares estimator omitting cluster g. As in Section 3.20, we can show that

$$\widetilde{\boldsymbol{e}}_g = \left(\boldsymbol{I}_{n_g} - \boldsymbol{X}_g\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\right)^{-1}\widehat{\boldsymbol{e}}_g \tag{4.57}$$

[5]This formula is due to Moulton (1990).

and

$$\widehat{\beta}_{(-g)} = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}_g'\widetilde{\boldsymbol{e}}_g. \tag{4.58}$$

We then have the robust covariance matrix estimator

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{CR3}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{g=1}^{G}\boldsymbol{X}_g'\widetilde{\boldsymbol{e}}_g\widetilde{\boldsymbol{e}}_g'\boldsymbol{X}_g\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{4.59}$$

The label "CR" refers to "cluster-robust" and "CR3" refers to the analogous formula for the HC3 estimator.

Similarly to the heteroskedastic-robust case you can show that CR3 is a conservative estimator for $\boldsymbol{V}_{\widehat{\beta}}$ in the sense that the conditional expectation of $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{CR3}}$ exceeds $\boldsymbol{V}_{\widehat{\beta}}$. This covariance matrix estimator is more cumbersome to implement, however, as the cluster-level prediction errors (4.57) cannot be calculated in a simple linear operation and requires a loop across clusters to calculate.

To illustrate in the context of the Kenyan schooling example we present the regression of student test scores on the school-level tracking dummy with two standard errors displayed. The first (in parenthesis) is the conventional robust standard error. The second [in square brackets] is the clustered standard error (4.55)-(4.56) where clustering is at the level of the school.

$$\begin{array}{rccc} \mathit{TestScore}_{ig} = - & 0.071 & + & 0.138 \quad \mathit{Tracking}_g + e_{ig}. \\ & (0.019) & & (0.026) \\ & [0.054] & & [0.078] \end{array} \tag{4.60}$$

We can see that the cluster-robust standard errors are roughly three times the conventional robust standard errors. Consequently, confidence intervals for the coefficients are greatly affected by the choice.

For illustration, we list here the commands needed to produce the regression results with clustered standard errors in Stata, R, and MATLAB.

Stata do File

```
*     Load data:
use "DDK2011.dta"
*     Standard the test score variable to have mean zero and unit variance:
egen testscore = std(totalscore)
*     Regression with standard errors clustered at the school level:
reg testscore tracking, cluster(schoolid)
```

You can see that clustered standard errors are simple to calculate in Stata.

R Program File

```
# Load the data and create variables
data <- read.table("DDK2011.txt",header=TRUE,sep="\ t")
y <- scale(as.matrix(data$totalscore))
n <- nrow(y)
x <- cbind(as.matrix(data$tracking),matrix(1,n,1))
schoolid <- as.matrix(data$schoolid)
k <- ncol(x)
xx <- t(x)%*%x
invx <- solve(xx)
beta <- solve(xx,t(x)%*%y)
xe <- x*rep(y-x%*%beta,times=k)
# Clustered robust standard error
xe_sum <- rowsum(xe,schoolid)
G <- nrow(xe_sum)
omega <- t(xe_sum)%*%xe_sum
scale <- G/(G-1)*(n-1)/(n-k)
V_clustered <- scale*invx%*%omega%*%invx
se_clustered <- sqrt(diag(V_clustered))
print(beta)
print(se_clustered)
```

Programming clustered standard errors in R is also relatively easy due to the convenient `rowsum` command which sums variables within clusters.

MATLAB Program File

```
% Load the data and create variables
data = xlsread('DDK2011.xlsx');
schoolid = data(:,2);
tracking = data(:,7);
totalscore = data(:,62);
y = (totalscore - mean(totalscore))./std(totalscore);
x = [tracking,ones(size(y,1),1)];
[n,k] = size(x);
xx = x'*x;
invx = inv(xx);
beta = xx\(x'*y)
e = y - x*beta;
% Clustered robust standard error
[schools,~,schoolidx] = unique(schoolid);
G = size(schools,1);
cluster_sums = zeros(G,k);
for j = 1:k
cluster_sums(:,j) = accumarray(schoolidx,x(:,j).*e);
end
omega = cluster_sums'*cluster_sums;
scale = G/(G-1)*(n-1)/(n-k);
V_clustered = scale*invx*omega*invx;
se_clustered = sqrt(diag(V_clustered));
display(beta);
display(se_clustered);
```

Here we see that programming clustered standard errors in MATLAB is less convenient than the other packages but still can be executed with just a few lines of code. This example uses the `accumarray` command which is similar to the `rowsum` command in R but only can be applied to vectors (hence the loop across the regressors) and works best if the *clusterid* variable are indices (which is why the original *schoolid* variable is transformed into indices in *schoolidx*. Application of these commands requires care and attention.

4.22 Inference with Clustered Samples

In this section we give some cautionary remarks and general advice about cluster-robust inference in econometric practice. There has been remarkably little theoretical research about the properties of cluster-robust methods – until quite recently – so these remarks may become dated rather quickly.

In many respects cluster-robust inference should be viewed similarly to heteroskedaticity-robust inference where a "cluster" in the cluster-robust case is interpreted similarly to an "observation" in the heteroskedasticity-robust case. In particular, the effective sample size should be viewed as the number of clusters, not the "sample size" n. This is because the cluster-robust covariance matrix estimator effectively treats each cluster as a single observation and estimates the covariance matrix based on the variation across cluster means. Hence if there are only $G = 50$ clusters inference should be viewed as

(at best) similar to heteroskedasticity-robust inference with $n = 50$ observations. This is a bit unsettling when the number of regressors is large (say $k = 20$) for then the covariance matrix will be estimated imprecisely.

Furthermore, most cluster-robust theory (for example, the work of Chris Hansen (2007)) assumes that the clusters are homogeneous including the assumption that the cluster sizes are all identical. This turns out to be a very important simplication. When this is violated – when, for example, cluster sizes are highly heterogeneous – the regression should be viewed as roughly equivalent to the heteroskedastic case with an extremely high degree of heteroskedasticity. Cluster sums have variances which are proportional to the cluster sizes so if the latter is heterogeneous so will be the variances of the cluster sums. This also has a large effect on finite sample inference. When clusters are heterogeneous then cluster-robust inference is similar to heteroskedasticity-robust inference with highly heteroskedastic observations.

Put together, if the number of clusters G is small and the number of observations per cluster is highly varied then we should interpret inferential statements with a great degree of caution. Unfortunately, small G with heterogeneous cluster sizes is commonplace. Many empirical studies on U.S. data cluster at the "state" level meaning that there are 50 or 51 clusters (the District of Columbia is typically treated as a state). The number of observations vary considerably across states since the populations are highly unequal. Thus when you read empirical papers with individual-level data but clustered at the "state" level you should be cautious and recognize that this is equivalent to inference with a small number of extremely heterogeneous observations.

A further complication occurs when we are interested in treatment as in the tracking example given in the previous section. In many cases (including Duflo, Dupas, and Kremer (2011)) the interest is in the effect of a treatment applied at the cluster level (e.g., schools). In many cases (not, however, Duflo, Dupas, and Kremer (2011)), the number of treated clusters is small relative to the total number of clusters; in an extreme case there is just a single treated cluster. Based on the reasoning given above these applications should be interpreted as equivalent to heteroskedasticity-robust inference with a sparse dummy variable as discussed in Section 4.16. As discussed there, standard error estimates can be erroneously small. In the extreme of a single treated cluster (in the example, if only a single school was tracked) then the estimated coefficient on *tracking* will be very imprecisely estimated yet will have a misleadingly small cluster standard error. In general, reported standard errors will greatly understate the imprecision of parameter estimates.

4.23 At What Level to Cluster?

A practical question which arises in the context of cluster-robust inference is "At what level should we cluster?" In some examples you could cluster at a very fine level, such as families or classrooms, or at higher levels of aggregation, such as neighborhoods, schools, towns, counties, or states. What is the correct level at which to cluster? Rules of thumb have been advocated by practitioners but at present there is little formal analysis to provide useful guidance. What do we know?

First, suppose cluster dependence is ignored or imposed at too fine a level (e.g. clustering by households instead of villages). Then variance estimators will be biased as they will omit covariance terms. As correlation is typically positive, this suggests that standard errors will be too small giving rise to spurious indications of significance and precision.

Second, suppose cluster dependence is imposed at too aggregate a measure (e.g. clustering by states rather than villages). This does not cause bias. But the variance estimators will contain many extra components so the precision of the covariance matrix estimator will be poor. This means that reported standard errors will be imprecise – more random – than if clustering had been less aggregate.

These considerations show that there is a trade-off between bias and variance in the estimation of the

covariance matrix by cluster-robust methods. It is not at all clear – based on current theory – what to do. I state this emphatically. We really do not know what is the "correct" level at which to do cluster-robust inference. This is a very interesting question and should certainly be explored by econometric research.

One challenge is that in empirical practice many people have observed: "Clustering is important. Standard errors change a lot whether or not we cluster. Therefore we should only report clustered standard errors." The flaw in this reasoning is that we do not know why in a specific empirical example the standard errors change under clustering. One possibility is that clustering reduces bias and thus is more accurate. The other possibility is that clustering adds sampling noise and is thus less accurate. In reality it is likely that both factors are present.

In any event a researcher should be aware of the number of clusters used in the reported calculations and should treat the number of clusters as the effective sample size for assessing inference. If the number of clusters is, say, $G = 20$, this should be treated as a very small sample.

To illustrate the thought experiment consider the empirical example of Duflo, Dupas, and Kremer (2011). They reported standard errors clustered at the school level and the application uses 111 schools. Thus $G = 111$ is the effective sample size. The number of observations (students) ranges from 19 to 62, which is reasonably homogeneous. This seems like a well balanced application of clustered variance estimation. However, one could imagine clustering at a different level of aggregation. We might consider clustering at a less aggregate level such as the classroom level, but this cannot be done in this particular application as there was only one classroom per school. Clustering at a more aggregate level could be done in this application at the level of the "zone". However, there are only 9 zones. Thus if we cluster by zone, $G = 9$ is the effective sample size which would lead to imprecise standard errors. In this particular example clustering at the school level (as done by the authors) is indeed the prudent choice.

4.24 Technical Proofs*

Proof of Theorems 4.4 and 4.5 Theorem 4.4 is a special case so we focus on Theorem 4.5. This argument is taken from B. E. Hansen (2021).

Our approach is to calculate the Cramér-Rao bound for a carefully crafted parametric model. This is based on an insight of Newey (1990, Appendix B) for the simpler context of a population expectation.

Without loss of generality, assume that the true coefficient equals $\beta_0 = 0$ and that $\sigma^2 = 1$. These are merely normalizations which simplify the notation. Also assume that $\boldsymbol{Y}$ has a joint density $f(\boldsymbol{y})$. This assumption can be avoided through use of the Radon-Nikodym derivative.

Define the truncation function $\mathbb{R}^n \to \mathbb{R}^n$

$$\psi_c(\boldsymbol{y}) = \boldsymbol{y}\mathbb{1}\left\{\left|\boldsymbol{X}'\Sigma^{-1}\boldsymbol{y}\right| \le c\right\} - \mathbb{E}\left[\boldsymbol{Y}\mathbb{1}\left\{\left|\boldsymbol{X}'\Sigma^{-1}\boldsymbol{Y}\right| \le c\right\}\right].$$

Notice that it satisfies $\left|\psi_c(\boldsymbol{y})\right| \le 2c$, $\mathbb{E}\left[\psi_c(\boldsymbol{Y})\right] = 0$, and

$$\mathbb{E}\left[\boldsymbol{Y}\psi_c(\boldsymbol{Y})'\right] = \mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}'\mathbb{1}\left\{\left|\boldsymbol{X}'\Sigma^{-1}\boldsymbol{Y}\right| \le c\right\}\right] \stackrel{\text{def}}{=} \Sigma_c.$$

As $c \to \infty$, $\Sigma_c \to \mathbb{E}\left[\boldsymbol{Y}\boldsymbol{Y}'\right] = \Sigma$. Pick c sufficiently large so that $\Sigma_c > 0$, which is feasible because $\Sigma > 0$.

Define the auxiliary joint density function

$$f_\beta\left(\boldsymbol{y}\right) = f(\boldsymbol{y})\left(1 + \psi_c(\boldsymbol{y})'\Sigma_c^{-1}\boldsymbol{X}\beta\right)$$

for parameters β in the set

$$B = \left\{\beta \in \mathbb{R}^m : \left\|\beta\right\| \le \frac{1}{2c}\right\}.$$

The bounds imply that for $\beta \in B$ and all $\boldsymbol{y}$

$$\left|\psi_c(\boldsymbol{y})'\Sigma_c^{-1}\boldsymbol{X}\beta\right| < 1.$$

This implies that f_β has the same support as f and satisfies the bounds

$$0 < f_\beta(\boldsymbol{y}) < 2f(\boldsymbol{y}). \tag{4.61}$$

We calculate that

$$\begin{aligned}\int f_\beta(\boldsymbol{y})\,d\boldsymbol{y} &= \int f(\boldsymbol{y})\,d\boldsymbol{y} + \int \psi_c(\boldsymbol{y})'\Sigma_c^{-1}\boldsymbol{X}\beta f_\beta(\boldsymbol{y})\,d\boldsymbol{y} \\ &= 1 + \mathbb{E}\left[\psi_c(\boldsymbol{Y})\right]'\Sigma_c^{-1}\boldsymbol{X}\beta \\ &= 1\end{aligned}$$

the last equality because $\mathbb{E}\left[\psi_c(\boldsymbol{Y})\right] = 0$. Together, these facts imply that f_β is a valid density function, and over $\beta \in B$ is a parametric family for $\boldsymbol{Y}$. Evaluated at $\beta_0 = 0$, which is in the interior of B, we see $f_0 = f$. This means that f_β is a correctly-specified parametric family with the true parameter value β_0.

Let $\mathbb{E}_\beta$ denote expectation under the density f_β. The expectation of $\boldsymbol{Y}$ in this model is

$$\begin{aligned}\mathbb{E}_\beta[\boldsymbol{Y}] &= \int \boldsymbol{y} f_\beta(\boldsymbol{y})\,d\boldsymbol{y} \\ &= \int \boldsymbol{y} f(\boldsymbol{y})\,d\boldsymbol{y} + \int \boldsymbol{y}\psi_c(\boldsymbol{y})'\Sigma_c^{-1}\boldsymbol{X}\beta f_\beta(\boldsymbol{y})\,d\boldsymbol{y} \\ &= \mathbb{E}[\boldsymbol{Y}] + \mathbb{E}\left[\boldsymbol{Y}\psi_c(\boldsymbol{Y})'\right]\Sigma_c^{-1}\boldsymbol{X}\beta \\ &= \boldsymbol{X}\beta\end{aligned}$$

because $\mathbb{E}[\boldsymbol{Y}] = 0$ and $\mathbb{E}\left[\boldsymbol{Y}\psi_c(\boldsymbol{Y})'\right] = \Sigma_c$. Thus, the model f_β is a linear regression with regression coefficient β.

The bound (4.61) implies

$$\mathbb{E}_\beta\left[\|\boldsymbol{Y}\|^2\right] = \int \|\boldsymbol{y}\|^2 f_\beta(\boldsymbol{y})\,d\boldsymbol{y} \le 2\int \|\boldsymbol{y}\|^2 f(\boldsymbol{y})\,d\boldsymbol{y} = 2\mathbb{E}\left[\|\boldsymbol{Y}\|^2\right] = 2\operatorname{tr}(\Sigma) < \infty.$$

This means that f_β has a finite variance for all $\beta \in B$.

The likelihood score for f_β is

$$\begin{aligned}S &= \frac{\partial}{\partial\beta}\log f_\beta(\boldsymbol{Y})\Big|_{\beta=0} \\ &= \frac{\partial}{\partial\beta}\log\left(1 + \psi_c(\boldsymbol{Y})'\Sigma_c^{-1}\boldsymbol{X}\beta\right)\Big|_{\beta=0} \\ &= \boldsymbol{X}'\Sigma_c^{-1}\psi_c(\boldsymbol{Y}).\end{aligned}$$

The information matrix is

$$\begin{aligned}\mathscr{I}_c &= \mathbb{E}\left[SS'\right] \\ &= \boldsymbol{X}'\Sigma_c^{-1}\mathbb{E}\left[\psi_c(\boldsymbol{Y})\psi_c(\boldsymbol{Y})'\right]\Sigma_c^{-1}\boldsymbol{X} \\ &\le \boldsymbol{X}'\Sigma_c^{-1}\boldsymbol{X},\end{aligned} \tag{4.62}$$

where the inequality is

$$\mathbb{E}\left[\psi_c(\boldsymbol{Y})\psi_c(\boldsymbol{Y})'\right] = \Sigma_c - \mathbb{E}\left[\boldsymbol{Y}\mathbb{1}\left\{\left|\boldsymbol{X}'\Sigma^{-1}\boldsymbol{Y}\right| \le c\right\}\right]\mathbb{E}\left[\boldsymbol{Y}\mathbb{1}\left\{\left|\boldsymbol{X}'\Sigma^{-1}\boldsymbol{Y}\right| \le c\right\}\right]' \le \Sigma_c.$$

By assumption, the estimator $\widetilde{\beta}$ is unbiased for β. The model f_β is regular (it is correctly specified as it contains the true density f, the support of $\boldsymbol{Y}$ does not depend on β, and the true value $\beta_0 = 0$ lies in the interior of B). Thus by the Cramér-Rao Theorem (Theorem 10.6 of *Probability and Statistics for Economists*)

$$\operatorname{var}\left[\widetilde{\beta}\right] \geq \mathscr{I}_c^{-1} \geq \left(\boldsymbol{X}'\Sigma_c^{-1}\boldsymbol{X}\right)^{-1}$$

where the second inequality is (4.62). Since this holds for all c, and $\Sigma_c \to \Sigma$ as $c \to \infty$,

$$\operatorname{var}\left[\widetilde{\beta}\right] \geq \limsup_{c\to\infty}\left(\boldsymbol{X}'\Sigma_c^{-1}\boldsymbol{X}\right)^{-1} = \left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}.$$

This is the variance lower bound. ■

4.25 Exercises

Exercise 4.1 For some integer k, set $\mu_k = \mathbb{E}[Y^k]$.

(a) Construct an estimator $\widehat{\mu}_k$ for μ_k.

(b) Show that $\widehat{\mu}_k$ is unbiased for μ_k.

(c) Calculate the variance of $\widehat{\mu}_k$, say $\operatorname{var}\left[\widehat{\mu}_k\right]$. What assumption is needed for $\operatorname{var}\left[\widehat{\mu}_k\right]$ to be finite?

(d) Propose an estimator of $\operatorname{var}\left[\widehat{\mu}_k\right]$.

Exercise 4.2 Calculate $\mathbb{E}\left[\left(\overline{Y} - \mu\right)^3\right]$, the skewness of $\overline{Y}$. Under what condition is it zero?

Exercise 4.3 Explain the difference between $\overline{Y}$ and μ. Explain the difference between $n^{-1}\sum_{i=1}^n X_i X_i'$ and $\mathbb{E}\left[X_i X_i'\right]$.

Exercise 4.4 True or False. If $Y = X'\beta + e$, $X \in \mathbb{R}$, $\mathbb{E}[e \mid X] = 0$, and $\widehat{e}_i$ is the OLS residual from the regression of Y_i on X_i, then $\sum_{i=1}^n X_i^2 \widehat{e}_i = 0$.

Exercise 4.5 Prove (4.20) and (4.21).

Exercise 4.6 Prove Theorem 4.5 under the restriction to linear estimators.

Exercise 4.7 Let $\widetilde{\beta}$ be the GLS estimator (4.22) under the assumptions (4.18) and (4.19). Assume that Σ is known and σ^2 is fdunknown. Define the residual vector $\widetilde{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\widetilde{\beta}$, and an estimator for σ^2

$$\widetilde{\sigma}^2 = \frac{1}{n-k}\widetilde{\boldsymbol{e}}'\Sigma^{-1}\widetilde{\boldsymbol{e}}.$$

(a) Show (4.23).

(b) Show (4.24).

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

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

(e) Find $\mathbb{E}\left[\widetilde{\sigma}^2 \mid \boldsymbol{X}\right]$.

(f) Is $\widetilde{\sigma}^2$ a reasonable estimator for σ^2?

Exercise 4.8 Let (Y_i, X_i) be a random sample with $\mathbb{E}[Y \mid X] = X'\beta$. Consider the **Weighted Least Squares** (WLS) estimator $\widetilde{\beta}_{\text{wls}} = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{Y}\right)$ where $\boldsymbol{W} = \text{diag}(w_1, ..., w_n)$ and $w_i = X_{ji}^{-2}$, where X_{ji} is one of the X_i.

(a) In which contexts would $\widetilde{\beta}_{\text{wls}}$ be a good estimator?

(b) Using your intuition, in which situations do you expect $\widetilde{\beta}_{\text{wls}}$ to perform better than OLS?

Exercise 4.9 Show (4.32) in the homoskedastic regression model.

Exercise 4.10 Prove (4.40).

Exercise 4.11 Show (4.41) in the homoskedastic regression model.

Exercise 4.12 Let $\mu = \mathbb{E}[Y]$, $\sigma^2 = \mathbb{E}\left[\left(Y-\mu\right)^2\right]$ and $\mu_3 = \mathbb{E}\left[\left(Y-\mu\right)^3\right]$ and consider the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Find $\mathbb{E}\left[\left(\overline{Y}-\mu\right)^3\right]$ as a function of μ, σ^2, μ_3 and n.

Exercise 4.13 Take the simple regression model $Y = X\beta + e$, $X \in \mathbb{R}$, $\mathbb{E}[e \mid X] = 0$. Define $\sigma_i^2 = \mathbb{E}\left[e_i^2 \mid X_i\right]$ and $\mu_{3i} = \mathbb{E}\left[e_i^3 \mid X_i\right]$ and consider the OLS coefficient $\widehat{\beta}$. Find $\mathbb{E}\left[\left(\widehat{\beta}-\beta\right)^3 \mid \boldsymbol{X}\right]$.

Exercise 4.14 Take a regression model $Y = X\beta + e$ with $\mathbb{E}[e \mid X] = 0$ and i.i.d. observations (Y_i, X_i) and scalar X. The parameter of interest is $\theta = \beta^2$. Consider the OLS estimators $\widehat{\beta}$ and $\widehat{\theta} = \widehat{\beta}^2$.

(a) Find $\mathbb{E}\left[\widehat{\theta} \mid \boldsymbol{X}\right]$ using our knowledge of $\mathbb{E}\left[\widehat{\beta} \mid \boldsymbol{X}\right]$ and $V_{\widehat{\beta}} = \text{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right]$. Is $\widehat{\theta}$ biased for θ?

(b) Suggest an (approximate) biased-corrected estimator $\widehat{\theta}^*$ using an estimator $\widehat{V}_{\widehat{\beta}}$ for $V_{\widehat{\beta}}$.

(c) For $\widehat{\theta}^*$ to be potentially unbiased, which estimator of $V_{\widehat{\beta}}$ is most appropriate?

Under which conditions is $\widehat{\theta}^*$ unbiased?

Exercise 4.15 Consider an i.i.d. sample $\{Y_i, X_i\}$ $i = 1, ..., n$ where X is $k \times 1$. Assume the linear conditional expectation model $Y = X'\beta + e$ with $\mathbb{E}[e \mid X] = 0$. Assume that $n^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}_k$ (orthonormal regressors). Consider the OLS estimator $\widehat{\beta}$.

(a) Find $\boldsymbol{V}_{\widehat{\beta}} = \text{var}\left[\widehat{\beta}\right]$

(b) In general, are $\widehat{\beta}_j$ and $\widehat{\beta}_\ell$ for $j \neq \ell$ correlated or uncorrelated?

(c) Find a sufficient condition so that $\widehat{\beta}_j$ and $\widehat{\beta}_\ell$ for $j \neq \ell$ are uncorrelated.

Exercise 4.16 Take the linear homoskedastic CEF

$$\begin{aligned} Y^* &= X'\beta + e \\ \mathbb{E}[e \mid X] &= 0 \\ \mathbb{E}\left[e^2 \mid X\right] &= \sigma^2 \end{aligned} \tag{4.63}$$

and suppose that Y^* is measured with error. Instead of Y^*, we observe $Y = Y^* + u$ where u is measurement error. Suppose that e and u are independent and

$$\begin{aligned}\mathbb{E}[u \mid X] &= 0\\ \mathbb{E}\left[u^2 \mid X\right] &= \sigma_u^2(X)\end{aligned}$$

(a) Derive an equation for Y as a function of X. Be explicit to write the error term as a function of the structural errors e and u. What is the effect of this measurement error on the model (4.63)?

(b) Describe the effect of this measurement error on OLS estimation of β in the feasible regression of the observed Y on X.

(c) Describe the effect (if any) of this measurement error on standard error calculation for $\widehat{\beta}$.

Exercise 4.17 Suppose that for the random variables (Y, X) with $X > 0$ an economic model implies

$$\mathbb{E}[Y \mid X] = \left(\gamma + \theta X\right)^{1/2}. \tag{4.64}$$

A friend suggests that you estimate γ and θ by the linear regression of Y^2 on X, that is, to estimate the equation

$$Y^2 = \alpha + \beta X + e. \tag{4.65}$$

(a) Investigate your friend's suggestion. Define $u = Y - \left(\gamma + \theta X\right)^{1/2}$. Show that $\mathbb{E}[u \mid X] = 0$ is implied by (4.64).

(b) Use $Y = \left(\gamma + \theta X\right)^{1/2} + u$ to calculate $\mathbb{E}\left[Y^2 \mid X\right]$. What does this tell you about the implied equation (4.65)?

(c) Can you recover either γ and/or θ from estimation of (4.65)? Are additional assumptions required?

(d) Is this a reasonable suggestion?

Exercise 4.18 Take the model

$$\begin{aligned}Y &= X_1'\beta_1 + X_2'\beta_2 + e\\ \mathbb{E}[e \mid X] &= 0\\ \mathbb{E}\left[e^2 \mid X\right] &= \sigma^2\end{aligned}$$

where $X = (X_1, X_2)$, with X_1 $k_1 \times 1$ and X_2 $k_2 \times 1$. Consider the short regression $Y_i = X_{1i}'\widehat{\beta}_1 + \widehat{e}_i$ and define the error variance estimator $s^2 = (n - k_1)^{-1}\sum_{i=1}^n \widehat{e}_i^2$. Find $\mathbb{E}\left[s^2 \mid \boldsymbol{X}\right]$.

Exercise 4.19 Let $\boldsymbol{Y}$ be $n \times 1$, $\boldsymbol{X}$ be $n \times k$, and $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{C}$ where $\boldsymbol{C}$ is $k \times k$ and full-rank. Let $\widehat{\beta}$ be the least squares estimator from the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$, and let $\widehat{\boldsymbol{V}}$ be the estimate of its asymptotic covariance matrix. Let $\widehat{\beta}^*$ and $\widehat{\boldsymbol{V}}^*$ be those from the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}^*$. Derive an expression for $\widehat{\boldsymbol{V}}^*$ as a function of $\widehat{\boldsymbol{V}}$.

Exercise 4.20 Take the model in vector notation

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X}\beta + \boldsymbol{e}\\ \mathbb{E}[\boldsymbol{e} \mid \boldsymbol{X}] &= 0\\ \mathbb{E}\left[\boldsymbol{e}\boldsymbol{e}' \mid \boldsymbol{X}\right] &= \Sigma.\end{aligned}$$

Assume for simplicity that Σ is known. Consider the OLS and GLS estimators $\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{Y}\right)$ and $\widetilde{\beta} = \left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{Y}\right)$. Compute the (conditional) covariance between $\widehat{\beta}$ and $\widetilde{\beta}$:

$$\mathbb{E}\left[\left(\widehat{\beta}-\beta\right)\left(\widetilde{\beta}-\beta\right)' \mid \boldsymbol{X}\right]$$

Find the (conditional) covariance matrix for $\widehat{\beta}-\widetilde{\beta}$:

$$\mathbb{E}\left[\left(\widehat{\beta}-\widetilde{\beta}\right)\left(\widehat{\beta}-\beta\right)' \mid \boldsymbol{X}\right].$$

Exercise 4.21 The model is

$$\begin{aligned} Y_i &= X_i'\beta + e_i \\ \mathbb{E}\left[e_i \mid X_i\right] &= 0 \\ \mathbb{E}\left[e_i^2 \mid X_i\right] &= \sigma_i^2 \\ \Sigma &= \operatorname{diag}\{\sigma_1^2, ..., \sigma_n^2\}. \end{aligned}$$

The parameter β is estimated by OLS $\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ and GLS $\widetilde{\beta} = \left(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\Sigma^{-1}\boldsymbol{Y}$. Let $\widehat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\beta}$ and $\widetilde{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\widetilde{\beta}$ denote the residuals. Let $\widehat{R}^2 = 1 - \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}}/(\boldsymbol{Y}^{*\prime}\boldsymbol{Y}^*)$ and $\widetilde{R}^2 = 1 - \widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}}/(\boldsymbol{Y}^{*\prime}\boldsymbol{Y}^*)$ denote the equation R^2 where $\boldsymbol{Y}^* = \boldsymbol{Y} - \overline{Y}$. If the error e_i is truly heteroskedastic will $\widehat{R}^2$ or $\widetilde{R}^2$ be smaller?

Exercise 4.22 An economist friend tells you that the assumption that the observations (Y_i, X_i) are i.i.d. implies that the regression $Y = X'\beta + e$ is homoskedastic. Do you agree with your friend? How would you explain your position?

Exercise 4.23 Take the linear regression model with $\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X}] = \boldsymbol{X}\beta$. Define the **ridge regression** estimator

$$\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{I}_k\lambda\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

where $\lambda > 0$ is a fixed constant. Find $E\left[\widehat{\beta} \mid \boldsymbol{X}\right]$. Is $\widehat{\beta}$ biased for β?

Exercise 4.24 Continue the empirical analysis in Exercise 3.24.

(a) Calculate standard errors using the homoskedasticity formula and using the four covariance matrices from Section 4.14.

(b) Repeat in a second programming language. Are they identical?

Exercise 4.25 Continue the empirical analysis in Exercise 3.26. Calculate standard errors using the HC3 method. Repeat in your second programming language. Are they identical?

Exercise 4.26 Extend the empirical analysis reported in Section 4.21 using the `DDK2011` dataset on the textbook website.. Do a regression of standardized test score (*totalscore* normalized to have zero mean and variance 1) on tracking, age, gender, being assigned to the contract teacher, and student's percentile in the initial distribution. (The sample size will be smaller as some observations have missing variables.) Calculate standard errors using both the conventional robust formula, and clustering based on the school.

(a) Compare the two sets of standard errors. Which standard error changes the most by clustering? Which changes the least?

(b) How does the coefficient on *tracking* change by inclusion of the individual controls (in comparison to the results from (4.60))?

Chapter 5

Normal Regression

5.1 Introduction

This chapter introduces the normal regression model, which is a special case of the linear regression model. It is important as normality allows precise distributional characterizations and sharp inferences. It also provides a baseline for comparison with alternative inference methods, such as asymptotic approximations and the bootstrap.

The normal regression model is a fully parametric setting where maximum likelihood estimation is appropriate. Therefore in this chapter we introduce likelihood methods. The method of maximum likelihood is a powerful statistical method for parametric models (such as the normal regression model) and is widely used in econometric practice.

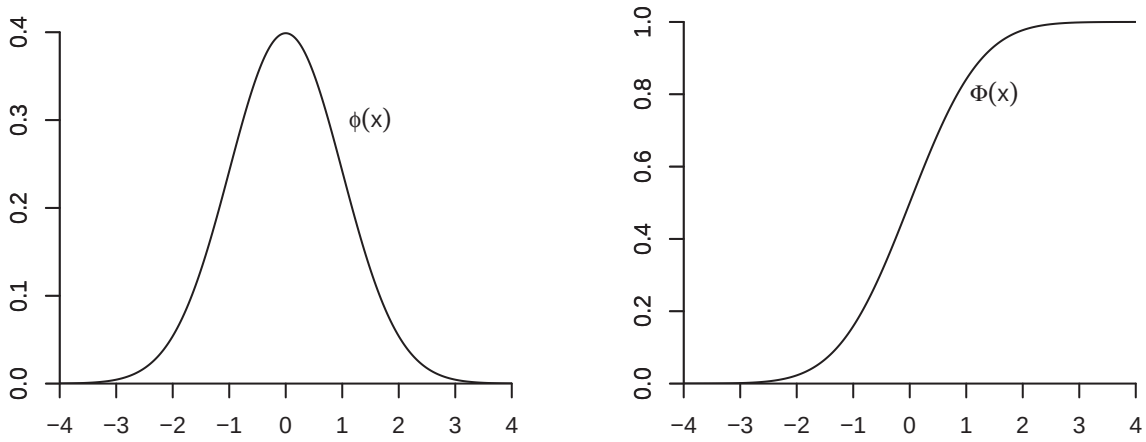
We start the chapter with a review of the definition and properties of the normal distribution. For detail and mathematical proofs see Chapter 5 of *Probability and Statistics for Economists*.

5.2 The Normal Distribution

We say that a random variable Z has the **standard normal distribution**, or **Gaussian**, written $Z \sim \mathrm{N}(0,1)$, if it has the density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad -\infty < x < \infty.$$

The standard normal density is typically written with the symbol $\phi(x)$ and the corresponding distribution function by $\Phi(x)$. Plots of the standard normal density function $\phi(x)$ and distribution function $\Phi(x)$ are displayed in Figure 5.1.



(a) Normal Density (b) Normal Distribution

Figure 5.1: Standard Normal Density and Distribution

Theorem 5.1 If $Z \sim \mathrm{N}(0,1)$ then

1. All integer moments of Z are finite.
2. All odd moments of Z equal 0.
3. For any positive integer m
$$\mathbb{E}\left[Z^{2m}\right] = (2m-1)!! = (2m-1) \times (2m-3) \times \cdots \times 1.$$
4. For any $r > 0$
$$\mathbb{E}|Z|^r = \frac{2^{r/2}}{\sqrt{\pi}}\Gamma\left(\frac{r+1}{2}\right)$$
where $\Gamma(t) = \int_0^\infty u^{t-1}e^{-u}du$ is the gamma function.

If $Z \sim \mathrm{N}(0,1)$ and $X = \mu + \sigma Z$ for $\mu \in \mathbb{R}$ and $\sigma \geq 0$ then X has the **univariate normal distribution**, written $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$. By change-of-variables X has the density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right), \qquad -\infty < x < \infty.$$

The expectation and variance of X are μ and σ^2, respectively.

The normal distribution and its relatives (the chi-square, student t, F, non-central chi-square, and F) are frequently used for inference to calculate critical values and p-values. This involves evaluating the normal cdf $\Phi(x)$ and its inverse. Since the cdf $\Phi(x)$ is not available in closed form, statistical textbooks

have traditionally provided tables for this purpose. Such tables are not used currently as these calculations are embedded in modern statistical software. For convenience, we list the appropriate commands in MATLAB, R, and Stata to compute the cumulative distribution function of commonly used statistical distributions.

Numerical Cumulative Distribution Functions

To calculate $\mathbb{P}[Z \leq x]$ for given x

	MATLAB	R	Stata
$\mathrm{N}(0,1)$	`normcdf(x)`	`pnorm(x)`	`normal(x)`
χ^2_r	`chi2cdf(x,r)`	`pchisq(x,r)`	`chi2(r,x)`
t_r	`tcdf(x,r)`	`pt(x,r)`	`1-ttail(r,x)`
$F_{r,k}$	`fcdf(x,r,k)`	`pf(x,r,k)`	`F(r,k,x)`
$\chi^2_r(d)$	`ncx2cdf(x,r,d)`	`pchisq(x,r,d)`	`nchi2(r,d,x)`
$F_{r,k}(d)$	`ncfcdf(x,r,k,d)`	`pf(x,r,k,d)`	`1-nFtail(r,k,d,x)`

Here we list the appropriate commands to compute the inverse probabilities (quantiles) of the same distributions.

Numerical Quantile Functions

To calculate x which solves $p = \mathbb{P}[Z \leq x]$ for given p

	MATLAB	R	Stata
$\mathrm{N}(0,1)$	`norminv(p)`	`qnorm(p)`	`invnormal(p)`
χ^2_r	`chi2inv(p,r)`	`qchisq(p,r)`	`invchi2(r,p)`
t_r	`tinv(p,r)`	`qt(p,r)`	`invttail(r,1-p)`
$F_{r,k}$	`finv(p,r,k)`	`qf(p,r,k)`	`invF(r,k,p)`
$\chi^2_r(d)$	`ncx2inv(p,r,d)`	`qchisq(p,r,d)`	`invnchi2(r,d,p)`
$F_{r,k}(d)$	`ncfinv(p,r,k,d)`	`qf(p,r,k,d)`	`invnFtail(r,k,d,1-p)`

5.3 Multivariate Normal Distribution

We say that the k-vector Z has a **multivariate standard normal distribution**, written $Z \sim \mathrm{N}(0, \boldsymbol{I}_k)$, if it has the joint density

$$f(x) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{x'x}{2}\right), \qquad x \in \mathbb{R}^k.$$

The mean and covariance matrix of Z are 0 and $\boldsymbol{I}_k$, respectively. The multivariate joint density factors into the product of univariate normal densities, so the elements of Z are mutually independent standard normals.

If $Z \sim \mathrm{N}(0, \boldsymbol{I}_k)$ and $X = \mu + \boldsymbol{B}Z$ then the k-vector X has a **multivariate normal distribution**, written $X \sim \mathrm{N}\left(\mu, \Sigma\right)$ where $\Sigma = \boldsymbol{B}\boldsymbol{B}' \geq 0$. If $\Sigma > 0$ then by change-of-variables X has the joint density function

$$f(x) = \frac{1}{(2\pi)^{k/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{\left(x-\mu\right)' \Sigma^{-1} \left(x-\mu\right)}{2}\right), \qquad x \in \mathbb{R}^k.$$

The expectation and covariance matrix of X are μ and Σ, respectively. By setting $k = 1$ you can check that the multivariate normal simplifies to the univariate normal.

An important property of normal random vectors is that affine functions are multivariate normal.

Theorem 5.2 If $X \sim \mathrm{N}\left(\mu, \Sigma\right)$ and $Y = \boldsymbol{a} + \boldsymbol{B}X$, then $Y \sim \mathrm{N}\left(\boldsymbol{a} + \boldsymbol{B}\mu, \boldsymbol{B}\Sigma\boldsymbol{B}'\right)$.

One simple implication of Theorem 5.2 is that if X is multivariate normal then each component of X is univariate normal.

Another useful property of the multivariate normal distribution is that uncorrelatedness is the same as independence. That is, if a vector is multivariate normal, subsets of variables are independent if and only if they are uncorrelated.

Theorem 5.3 Properties of the Multivariate Normal Distribution

1. The expectation and covariance matrix of $X \sim \mathrm{N}\left(\mu, \Sigma\right)$ are $\mathbb{E}[X] = \mu$ and $\mathrm{var}[X] = \Sigma$.
2. If (X, Y) are multivariate normal, X and Y are uncorrelated if and only if they are independent.
3. If $X \sim \mathrm{N}\left(\mu, \Sigma\right)$ and $Y = \boldsymbol{a} + \boldsymbol{B}X$, then $Y \sim \mathrm{N}\left(\boldsymbol{a} + \boldsymbol{B}\mu, \boldsymbol{B}\Sigma\boldsymbol{B}'\right)$.
4. If $X \sim \mathrm{N}(0, \boldsymbol{I}_k)$ then $X'X \sim \chi^2_k$, chi-square with k degrees of freedom.
5. If $X \sim \mathrm{N}(0, \Sigma)$ with $\Sigma > 0$ then $X'\Sigma^{-1}X \sim \chi^2_k$ where $k = \dim(X)$.
6. If $X \sim \mathrm{N}(\mu, \Sigma)$ with $\Sigma > 0$, $r \times r$, then $X'\Sigma^{-1}X \sim \chi^2_r(\lambda)$ where $\lambda = \mu'\Sigma^{-1}\mu$.
7. If $Z \sim \mathrm{N}(0,1)$ and $Q \sim \chi^2_k$ are independent then $Z/\sqrt{Q/k} \sim t_k$, student t with k degrees of freedom.
8. If (Y, X) are multivariate normal

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix}\right)$$

with $\Sigma_{YY} > 0$ and $\Sigma_{XX} > 0$, then the conditional distributions are

$$\begin{aligned} Y \mid X &\sim \mathrm{N}\left(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}\left(X - \mu_X\right), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right) \\ X \mid Y &\sim \mathrm{N}\left(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}\left(Y - \mu_Y\right), \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\right). \end{aligned}$$

5.4 Joint Normality and Linear Regression

Suppose the variables (Y, X) are jointly normally distributed. Consider the best linear predictor of Y given X

$$Y = X'\beta + \alpha + e.$$

By the properties of the best linear predictor, $\mathbb{E}[Xe] = 0$ and $\mathbb{E}[e] = 0$, so X and e are uncorrelated. Since (e, X) is an affine transformation of the normal vector (Y, X) it follows that (e, X) is jointly normal (Theorem 5.2). Since (e, X) is jointly normal and uncorrelated they are independent (Theorem 5.3). Independence implies that

$$\mathbb{E}[e \mid X] = \mathbb{E}[e] = 0$$

and

$$\mathbb{E}\left[e^2 \mid X\right] = \mathbb{E}\left[e^2\right] = \sigma^2$$

which are properties of a homoskedastic linear CEF.

We have shown that when (Y, X) are jointly normally distributed they satisfy a normal linear CEF

$$Y = X'\beta + \alpha + e$$

where

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

is independent of X. This result can also be deduced from Theorem 5.3.7.

This is a classical motivation for the linear regression model.

5.5 Normal Regression Model

The normal regression model is the linear regression model with an independent normal error

$$\begin{aligned} Y &= X'\beta + e \\ e &\sim \mathrm{N}(0, \sigma^2). \end{aligned} \tag{5.1}$$

As we learned in Section 5.4, the normal regression model holds when (Y, X) are jointly normally distributed. Normal regression, however, does not require joint normality. All that is required is that the conditional distribution of Y given X is normal (the marginal distribution of X is unrestricted). In this sense the normal regression model is broader than joint normality. Notice that for notational convenience we have written (5.1) so that X contains the intercept.

Normal regression is a parametric model where likelihood methods can be used for estimation, testing, and distribution theory. The **likelihood** is the name for the joint probability density of the data, evaluated at the observed sample, and viewed as a function of the parameters. The maximum likelihood estimator is the value which maximizes this likelihood function. Let us now derive the likelihood of the normal regression model.

First, observe that model (5.1) is equivalent to the statement that the conditional density of Y given X takes the form

$$f\left(y \mid x\right) = \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y - x'\beta\right)^2\right).$$

Under the assumption that the observations are mutually independent this implies that the conditional density of $(Y_1, ..., Y_n)$ given $(X_1, ..., X_n)$ is

$$\begin{aligned} f\left(y_1, ..., y_n \mid x_1, ..., x_n\right) &= \prod_{i=1}^{n} f\left(y_i \mid x_i\right) \\ &= \prod_{i=1}^{n} \frac{1}{\left(2\pi\sigma^2\right)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i - x_i'\beta\right)^2\right) \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2\right) \\ &\stackrel{\text{def}}{=} L_n(\beta, \sigma^2). \end{aligned}$$

This is called the **likelihood function** when evaluated at the sample data.

For convenience it is typical to work with the natural logarithm

$$\log L_n(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2 \stackrel{\text{def}}{=} \ell_n(\beta, \sigma^2) \tag{5.2}$$

which is called the **log-likelihood function**.

The **maximum likelihood estimator** (**MLE**) $(\widehat{\beta}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}})$ is the value which maximizes the log-likelihood. We can write the maximization problem as

$$(\widehat{\beta}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}}) = \underset{\beta\in\mathbb{R}^k,\, \sigma^2>0}{\operatorname{argmax}}\ \ell_n(\beta, \sigma^2). \tag{5.3}$$

In most applications of maximum likelihood the MLE must be found by numerical methods. However in the case of the normal regression model we can find an explicit expression for $\widehat{\beta}_{\text{mle}}$ and $\widehat{\sigma}^2_{\text{mle}}$.

The maximizers $(\widehat{\beta}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}})$ of (5.3) jointly solve the first-order conditions (FOC)

$$0 = \frac{\partial}{\partial\beta}\ell_n(\beta, \sigma^2)\Big|_{\beta=\widehat{\beta}_{\text{mle}}, \sigma^2=\widehat{\sigma}^2_{\text{mle}}} = \frac{1}{\widehat{\sigma}^2_{\text{mle}}}\sum_{i=1}^{n} X_i\left(Y_i - X_i'\widehat{\beta}_{\text{mle}}\right) \tag{5.4}$$

$$0 = \frac{\partial}{\partial\sigma^2}\ell_n(\beta, \sigma^2)\Big|_{\beta=\widehat{\beta}_{\text{mle}}, \sigma^2=\widehat{\sigma}^2_{\text{mle}}} = -\frac{n}{2\widehat{\sigma}^2_{\text{mle}}} + \frac{1}{2\widehat{\sigma}^4_{\text{mle}}}\sum_{i=1}^{n}\left(Y_i - X_i'\widehat{\beta}_{\text{mle}}\right)^2. \tag{5.5}$$

The first FOC (5.4) is proportional to the first-order conditions for the least squares minimization problem of Section 3.6. It follows that the MLE satisfies

$$\widehat{\beta}_{\text{mle}} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right) = \widehat{\beta}_{\text{ols}}.$$

That is, the MLE for β is algebraically identical to the OLS estimator.

Solving the second FOC (5.5) for $\widehat{\sigma}^2_{\text{mle}}$ we find

$$\widehat{\sigma}^2_{\text{mle}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\widehat{\beta}_{\text{mle}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\widehat{\beta}_{\text{ols}}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2 = \widehat{\sigma}^2_{\text{ols}}.$$

Thus the MLE for σ^2 is identical to the OLS/moment estimator from (3.26).

Since the OLS estimator and MLE under normality are equivalent, $\widehat{\beta}$ is described by some authors as the maximum likelihood estimator, and by other authors as the least squares estimator. It is important

to remember, however, that $\widehat{\beta}$ is only the MLE when the error e has a known normal distribution and not otherwise.

Plugging the estimators into (5.2) we obtain the maximized log-likelihood

$$\ell_n\left(\widehat{\beta}_{\text{mle}}, \widehat{\sigma}^2_{\text{mle}}\right) = -\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2_{\text{mle}}\right) - \frac{n}{2}. \tag{5.6}$$

The log-likelihood is typically reported as a measure of fit.

It may seem surprising that the MLE $\widehat{\beta}_{\text{mle}}$ is algebraically 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

Carl Friedrich Gauss (1777-1855) was one of the most influential mathematicians in history. His contributions impact many topics of importance to economics and econometrics, including the Gauss-Markov Theorem, the Gauss-Newton algorithm, and Gaussian elimination. In a 1809 paper, he set the regression model on probabilistic foundations by proposing that the equation errors be treated as random variables. He showed that if the error distribution takes the form we now call normal (or "Gaussian") then the MLE for the coefficients equals the least squares estimator.

5.6 Distribution of OLS Coefficient Vector

In the normal linear regression model we can derive exact sampling distributions for the OLS/MLE estimator, residuals, and variance estimator. In this section we derive the distribution of the OLS coefficient estimator.

The normality assumption $e \mid X \sim \mathrm{N}\left(0, \sigma^2\right)$ combined with independence of the observations has the multivariate implication

$$\boldsymbol{e} \mid \boldsymbol{X} \sim \mathrm{N}\left(0, \boldsymbol{I}_n\sigma^2\right).$$

That is, the error vector $\boldsymbol{e}$ is independent of $\boldsymbol{X}$ and is normally distributed.

Recall that the OLS estimator satisfies

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

which is a linear function of $\boldsymbol{e}$. Since linear functions of normals are also normal (Theorem 5.2) this implies that conditional on $\boldsymbol{X}$,

$$\begin{aligned}
\widehat{\beta} - \beta \mid \boldsymbol{X} &\sim \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathrm{N}\left(0, \boldsymbol{I}_n\sigma^2\right) \\
&\sim \mathrm{N}\left(0, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) \\
&= \mathrm{N}\left(0, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).
\end{aligned}$$

This shows that under the assumption of normal errors the OLS estimator has an exact normal distribution.

Theorem 5.4 In the normal regression model,

$$\widehat{\beta} \mid \boldsymbol{X} \sim \mathrm{N}\left(\beta, \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right).$$

Theorems 5.2 and 5.4 imply that any affine function of the OLS estimator is also normally distributed including individual components. Letting β_j and $\widehat{\beta}_j$ denote the j^{th} elements of β and $\widehat{\beta}$, we have

$$\widehat{\beta}_j \mid \boldsymbol{X} \sim \mathrm{N}\left(\beta_j, \sigma^2 \left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}\right). \tag{5.7}$$

Theorem 5.4 is a statement about the conditional distribution. What about the unconditional distribution? In Section 4.7 we presented Kinal's theorem about the existence of moments for the joint normal regression model. We re-state the result here.

Theorem 5.5 Kinal (1980) If (Y, X) are jointly normal, then for any r, $\mathbb{E}\left\|\widehat{\beta}\right\|^r < \infty$ if and only if $r < n - k + 1$.

5.7 Distribution of OLS Residual Vector

Consider the OLS residual vector. Recall from (3.24) that $\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. This shows that $\widehat{\boldsymbol{e}}$ is linear in $\boldsymbol{e}$. So conditional on $\boldsymbol{X}$

$$\widehat{\boldsymbol{e}} = \boldsymbol{M}\boldsymbol{e} \mid \boldsymbol{X} \sim \mathrm{N}\left(0, \sigma^2 \boldsymbol{M}\boldsymbol{M}\right) = \mathrm{N}\left(0, \sigma^2 \boldsymbol{M}\right)$$

the final equality because $\boldsymbol{M}$ is idempotent (see Section 3.12). This shows that the residual vector has an exact normal distribution.

Furthermore, it is useful to find the joint distribution of $\widehat{\beta}$ and $\widehat{\boldsymbol{e}}$. This is easiest done by writing the two as a stacked linear function of the error $\boldsymbol{e}$. Indeed,

$$\begin{pmatrix} \widehat{\beta} - \beta \\ \widehat{\boldsymbol{e}} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e} \\ \boldsymbol{M}\boldsymbol{e} \end{pmatrix} = \begin{pmatrix} \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' \\ \boldsymbol{M} \end{pmatrix} \boldsymbol{e}$$

which is a linear function of $\boldsymbol{e}$. The vector has a joint normal distribution with covariance matrix

$$\begin{pmatrix} \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} & 0 \\ 0 & \sigma^2 \boldsymbol{M} \end{pmatrix}.$$

The off-diagonal block is zero because $\boldsymbol{X}'\boldsymbol{M} = 0$ from (3.21). Since this is zero it follows that $\widehat{\beta}$ and $\widehat{\boldsymbol{e}}$ are statistically independent (Theorem 5.3.2).

Theorem 5.6 In the normal regression model, $\widehat{\boldsymbol{e}} \mid \boldsymbol{X} \sim \mathrm{N}\left(0, \sigma^2 \boldsymbol{M}\right)$ and is independent of $\widehat{\beta}$.

The fact that $\widehat{\beta}$ and $\widehat{\boldsymbol{e}}$ are independent implies that $\widehat{\beta}$ is independent of any function of the residual vector including individual residuals $\widehat{e}_i$ and the variance estimators s^2 and $\widehat{\sigma}^2$.

5.8 Distribution of Variance Estimator

Next, consider the variance estimator s^2 from (4.31). Using (3.28) it satisfies $(n-k)\, s^2 = \widehat{\boldsymbol{e}}'\widehat{\boldsymbol{e}} = \boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}$. The spectral decomposition of $\boldsymbol{M}$ (equation (A.4)) is $\boldsymbol{M} = \boldsymbol{H}\Lambda\boldsymbol{H}'$ where $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$ and Λ is diagonal with the eigenvalues of $\boldsymbol{M}$ on the diagonal. Since $\boldsymbol{M}$ is idempotent with rank $n-k$ (see Section 3.12) it has $n-k$ eigenvalues equalling 1 and k eigenvalues equalling 0, so

$$\Lambda = \begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}_k \end{bmatrix}.$$

Let $\boldsymbol{u} = \boldsymbol{H}'\boldsymbol{e} \sim \mathrm{N}\left(\boldsymbol{0}, \boldsymbol{I}_n\sigma^2\right)$ (see Exercise 5.2) and partition $\boldsymbol{u} = \left(\boldsymbol{u}_1', \boldsymbol{u}_2'\right)'$ where $\boldsymbol{u}_1 \sim \mathrm{N}\left(0, \boldsymbol{I}_{n-k}\sigma^2\right)$. Then

$$\begin{aligned}
(n-k)\, s^2 &= \boldsymbol{e}'\boldsymbol{M}\boldsymbol{e} \\
&= \boldsymbol{e}'\boldsymbol{H}\begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{H}'\boldsymbol{e} \\
&= \boldsymbol{u}'\begin{bmatrix} \boldsymbol{I}_{n-k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}\boldsymbol{u} \\
&= \boldsymbol{u}_1'\boldsymbol{u}_1 \\
&\sim \sigma^2\chi^2_{n-k}.
\end{aligned}$$

We see that in the normal regression model the exact distribution of s^2 is a scaled chi-square.

Since $\widehat{\boldsymbol{e}}$ is independent of $\widehat{\beta}$ it follows that s^2 is independent of $\widehat{\beta}$ as well.

Theorem 5.7 In the normal regression model,

$$\frac{(n-k)\, s^2}{\sigma^2} \sim \chi^2_{n-k}$$

and is independent of $\widehat{\beta}$.

5.9 t-statistic

An alternative way of writing (5.7) is

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right]_{jj}}} \sim \mathrm{N}(0,1).$$

This is sometimes called a **standardized** statistic as the distribution is the standard normal.

Now take the standardized statistic and replace the unknown variance σ^2 with its estimator s^2. We call this a **t-ratio** or **t-statistic**

$$T = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{s^2 \left[\left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \right]_{jj}}} = \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)}$$

where $s(\widehat{\beta}_j)$ is the classical (homoskedastic) standard error for $\widehat{\beta}_j$ from (4.42). We will sometimes write the t-statistic as $T(\beta_j)$ to explicitly indicate its dependence on the parameter value β_j, and sometimes will simplify notation and write the t-statistic as T when the dependence is clear from the context.

With algebraic re-scaling we can write the t-statistic as the ratio of the standardized statistic and the square root of the scaled variance estimator. Since the distributions of these two components are normal and chi-square, respectively, and independent, we deduce that the t-statistic has the distribution

$$\begin{aligned} T &= \frac{\widehat{\beta}_j - \beta_j}{\sqrt{\sigma^2 \left[\left(\boldsymbol{X}'\boldsymbol{X} \right)^{-1} \right]_{jj}}} \Bigg/ \sqrt{\frac{(n-k)s^2}{\sigma^2} \Bigg/ (n-k)} \\ &\sim \frac{\mathrm{N}(0,1)}{\sqrt{\chi^2_{n-k}/(n-k)}} \\ &\sim t_{n-k} \end{aligned}$$

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

This derivation shows that the t-ratio has a sampling distribution which depends only on the quantity $n-k$. The distribution does not depend on any other features of the data. In this context, we say that the distribution of the t-ratio is **pivotal**, meaning that it does not depend on unknowns.

The trick behind this result is scaling the centered coefficient by its standard error, and recognizing that each depends on the unknown σ only through scale. Thus the ratio of the two does not depend on σ. This trick (scaling to eliminate dependence on unknowns) is known as **studentization**.

Theorem 5.8 In the normal regression model, $T \sim t_{n-k}$.

An important caveat about Theorem 5.8 is that it only applies to the t-statistic constructed with the homoskedastic (old-fashioned) standard error. It does not apply to a t-statistic constructed with any of the robust standard errors. In fact, the robust t-statistics can have finite sample distributions which deviate considerably from t_{n-k} even when the regression errors are independent $\mathrm{N}(0,\sigma^2)$. Thus the distributional result in Theorem 5.8 and the use of the t distribution in finite samples is only exact when applied to classical t-statistics under the normality assumption.

5.10 Confidence Intervals for Regression Coefficients

The OLS estimator $\widehat{\beta}$ is a **point estimator** for a coefficient β. A broader concept is a **set** or **interval estimator** which takes the form $\widehat{C} = [\widehat{L}, \widehat{U}]$. The goal of an interval estimator $\widehat{C}$ is to contain the true value, e.g. $\beta \in \widehat{C}$, with high probability.

The interval estimator $\widehat{C}$ is a function of the data and hence is random.

An interval estimator $\widehat{C}$ is called a $1-\alpha$ **confidence interval** when $\mathbb{P}\left[\beta \in \widehat{C}\right] = 1-\alpha$ for a selected value of α. The value $1-\alpha$ is called the **coverage probability**. Typical choices for the coverage probability $1-\alpha$ are 0.95 or 0.90.

The probability calculation $\mathbb{P}\left[\beta \in \widehat{C}\right]$ is easily mis-interpreted as treating β as random and $\widehat{C}$ as fixed. (The probability that β is in $\widehat{C}$.) This is not the appropriate interpretation. Instead, the correct interpretation is that the probability $\mathbb{P}\left[\beta \in \widehat{C}\right]$ treats the point β as fixed and the set $\widehat{C}$ as random. It is the probability that the random set $\widehat{C}$ covers (or contains) the fixed true coefficient β.

There is not a unique method to construct confidence intervals. For example, one simple (yet silly) interval is

$$\widehat{C} = \begin{cases} \mathbb{R} & \text{with probability } 1-\alpha \\ \{\widehat{\beta}\} & \text{with probability } \alpha. \end{cases}$$

If $\widehat{\beta}$ has a continuous distribution, then by construction $\mathbb{P}\left[\beta \in \widehat{C}\right] = 1-\alpha$, so this confidence interval has perfect coverage. However, $\widehat{C}$ is uninformative about $\widehat{\beta}$ and is therefore not useful.

Instead, a good choice for a confidence interval for the regression coefficient β is obtained by adding and subtracting from the estimator $\widehat{\beta}$ a fixed multiple of its standard error:

$$\widehat{C} = \left[\widehat{\beta} - c \times s(\widehat{\beta}),\quad \widehat{\beta} + c \times s(\widehat{\beta})\right] \tag{5.8}$$

where $c > 0$ is a pre-specified constant. This confidence interval is symmetric about the point estimator $\widehat{\beta}$ and its length is proportional to the standard error $s(\widehat{\beta})$.

Equivalently, $\widehat{C}$ is the set of parameter values for β such that the t-statistic $T(\beta)$ is smaller (in absolute value) than c, that is

$$\widehat{C} = \left\{\beta : \left|T(\beta)\right| \le c\right\} = \left\{\beta : -c \le \frac{\widehat{\beta}-\beta}{s(\widehat{\beta})} \le c\right\}.$$

The coverage probability of this confidence interval is

$$\begin{aligned} \mathbb{P}\left[\beta \in \widehat{C}\right] &= \mathbb{P}\left[\left|T(\beta)\right| \le c\right] \\ &= \mathbb{P}\left[-c \le T(\beta) \le c\right]. \end{aligned} \tag{5.9}$$

Since the t-statistic $T(\beta)$ has the t_{n-k} distribution, (5.9) equals $F(c) - F(-c)$, where $F(u)$ is the student t distribution function with $n-k$ degrees of freedom. Since $F(-c) = 1 - F(c)$ (see Exercise 5.8), we can write (5.9) as

$$\mathbb{P}\left[\beta \in \widehat{C}\right] = 2F(c) - 1.$$

This is the **coverage probability** of the interval $\widehat{C}$, and only depends on the constant c.

As we mentioned before, a confidence interval has the coverage probability $1-\alpha$. This requires selecting the constant c so that $F(c) = 1-\alpha/2$. This holds if c equals the $1-\alpha/2$ quantile of the t_{n-k} distribution. As there is no closed form expression for these quantiles we compute their values numerically. For example, by `tinv(1-alpha/2,n-k)` in MATLAB. With this choice the confidence interval (5.8) has exact coverage probability $1-\alpha$. By default, Stata reports 95% confidence intervals $\widehat{C}$ for each estimated regression coefficient using the same formula.

Theorem 5.9 In the normal regression model, (5.8) with $c = F^{-1}(1-\alpha/2)$ has coverage probability $\mathbb{P}\left[\beta \in \widehat{C}\right] = 1-\alpha$.

When the degree of freedom is large the distinction between the student t and the normal distribution is negligible. In particular, for $n-k \geq 61$ we have $c \leq 2.00$ for a 95% interval. Using this value we obtain the most commonly used confidence interval in applied econometric practice:

$$\widehat{C} = \left[\widehat{\beta} - 2s(\widehat{\beta}),\quad \widehat{\beta} + 2s(\widehat{\beta})\right]. \tag{5.10}$$

This is a useful rule-of-thumb. This 95% confidence interval $\widehat{C}$ is simple to compute and can be easily calculated from coefficient estimates and standard errors.

Theorem 5.10 In the normal regression model, if $n-k \geq 61$ then (5.10) has coverage probability $\mathbb{P}\left[\beta \in \widehat{C}\right] \geq 0.95$.

Confidence intervals are a simple yet effective tool to assess estimation uncertainty. When reading a set of empirical results look at the estimated coefficient estimates and the standard errors. For a parameter of interest compute the confidence interval $\widehat{C}$ and consider the meaning of the spread of the suggested values. If the range of values in the confidence interval are too wide to learn about β then do not jump to a conclusion about β based on the point estimate alone.

5.11 Confidence Intervals for Error Variance

We can also construct a confidence interval for the regression error variance σ^2 using the sampling distribution of s^2 from Theorem 5.7. This states that in the normal regression model

$$\frac{(n-k)\,s^2}{\sigma^2} \sim \chi^2_{n-k}. \tag{5.11}$$

Let $F(u)$ denote the χ^2_{n-k} distribution function and for some α set $c_1 = F^{-1}(\alpha/2)$ and $c_2 = F^{-1}(1-\alpha/2)$ (the $\alpha/2$ and $1-\alpha/2$ quantiles of the χ^2_{n-k} distribution). Equation (5.11) implies that

$$\mathbb{P}\left[c_1 \leq \frac{(n-k)\,s^2}{\sigma^2} \leq c_2\right] = F(c_2) - F(c_1) = 1-\alpha.$$

Rewriting the inequalities we find

$$\mathbb{P}\left[\frac{(n-k)\,s^2}{c_2} \leq \sigma^2 \leq \frac{(n-k)\,s^2}{c_1}\right] = 1-\alpha.$$

This shows that an exact $1-\alpha$ confidence interval for σ^2 is

$$\widehat{C} = \left[\frac{(n-k)\,s^2}{c_2},\quad \frac{(n-k)\,s^2}{c_1}\right]. \tag{5.12}$$

Theorem 5.11 In the normal regression model (5.12) has coverage probability $\mathbb{P}\left[\sigma^2 \in \widehat{C}\right] = 1-\alpha$.

The confidence interval (5.12) for σ^2 is asymmetric about the point estimate s^2 due to the latter's asymmetric sampling distribution.

5.12 t Test

A typical goal in an econometric exercise is to assess whether or not a coefficient β equals a specific value β_0. Often the specific value to be tested is $\beta_0 = 0$ but this is not essential. This is called **hypothesis testing**, a subject which will be explored in detail in Chapter 9. In this section and the following we give a short introduction specific to the normal regression model.

For simplicity write the coefficient to be tested as β. The **null hypothesis** is

$$\mathbb{H}_0 : \beta = \beta_0. \tag{5.13}$$

This states that the hypothesis is that the true value of β equals the hypothesized value β_0.

The alternative hypothesis is the complement of $\mathbb{H}_0$, and is written as

$$\mathbb{H}_1 : \beta \neq \beta_0.$$

This states that the true value of β does not equal the hypothesized value.

We are interested in testing $\mathbb{H}_0$ against $\mathbb{H}_1$. The method is to design a statistic which is informative about $\mathbb{H}_1$. If the observed value of the statistic is consistent with random variation under the assumption that $\mathbb{H}_0$ is true, then we deduce that there is no evidence against $\mathbb{H}_0$ and consequently do not reject $\mathbb{H}_0$. However, if the statistic takes a value which is unlikely to occur under the assumption that $\mathbb{H}_0$ is true, then we deduce that there is evidence against $\mathbb{H}_0$ and consequently we reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$. The main steps are to design a test statistic and to characterize its sampling distribution.

The standard statistic to test $\mathbb{H}_0$ against $\mathbb{H}_1$ is the absolute value of the t-statistic

$$|T| = \left| \frac{\widehat{\beta} - \beta_0}{s(\widehat{\beta})} \right|. \tag{5.14}$$

If $\mathbb{H}_0$ is true then we expect $|T|$ to be small, but if $\mathbb{H}_1$ is true then we would expect $|T|$ to be large. Hence the standard rule is to reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ for large values of the t-statistic $|T|$ and otherwise fail to reject $\mathbb{H}_0$. Thus the hypothesis test takes the form

$$\text{Reject } \mathbb{H}_0 \text{ if } |T| > c.$$

The constant c which appears in the statement of the test is called the **critical value**. Its value is selected to control the probability of false rejections. When the null hypothesis is true T has an exact t_{n-k} distribution in the normal regression model. Thus for a given value of c the probability of false rejection is

$$\begin{aligned}
\mathbb{P}\left[\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_0\right] &= \mathbb{P}\left[|T| > c \mid \mathbb{H}_0\right] \\
&= \mathbb{P}\left[T > c \mid \mathbb{H}_0\right] + \mathbb{P}\left[T < -c \mid \mathbb{H}_0\right] \\
&= 1 - F(c) + F(-c) \\
&= 2(1 - F(c))
\end{aligned}$$

where $F(u)$ is the t_{n-k} distribution function. This is the probability of false rejection and is decreasing in the critical value c. We select the value c so that this probability equals a pre-selected value called the **significance level** which is typically written as α. It is conventional to set $\alpha = 0.05$, though this is not a hard rule. We then select c so that $F(c) = 1 - \alpha/2$, which means that c is the $1 - \alpha/2$ quantile (inverse CDF) of the t_{n-k} distribution, the same as used for confidence intervals. With this choice the decision rule "Reject $\mathbb{H}_0$ if $|T| > c$" has a significance level (false rejection probability) of α.

Theorem 5.12 In the normal regression model if the null hypothesis (5.13) is true, then for $|T|$ defined in (5.14) $T \sim t_{n-k}$. If c is set so that $\mathbb{P}[|t_{n-k}| \geq c] = \alpha$,then the test "Reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $|T| > c$" has significance level α.

To report the result of a hypothesis test we need to pre-determine the significance level α in order to calculate the critical value c. This can be inconvenient and arbitrary. A simplification is to report what is known as the **p-value** of the test. In general, when a test takes the form "Reject $\mathbb{H}_0$ if $S > c$" and S has null distribution $G(u)$ then the p-value of the test is $p = 1 - G(S)$. A test with significance level α can be restated as "Reject $\mathbb{H}_0$ if $p < \alpha$". It is sufficient to report the p-value p and we can interpret the value of p as indexing the test's strength of rejection of the null hypothesis. Thus a p-value of 0.07 might be interpreted as "nearly significant", 0.05 as "borderline significant", and 0.001 as "highly significant". In the context of the normal regression model the p-value of a t-statistic $|T|$ is $p = 2(1 - F_{n-k}(|T|))$ where F_{n-k} is the t_{n-k} CDF. For example, in MATLAB the calculation is `2*(1-tcdf(abs(t),n-k))`. In Stata, the default is that for any estimated regression, t-statistics for each estimated coefficient are reported along with their p-values calculated using this same formula. These t-statistics test the hypotheses that each coefficient is zero.

A p-value reports the strength of evidence against $\mathbb{H}_0$ but is not itself a probability. A common misunderstanding is that the p-value is the "probability that the null hypothesis is true". This is an incorrect interpretation. It is a statistic, is random, and is a measure of the evidence against $\mathbb{H}_0$. Nothing more.

5.13 Likelihood Ratio Test

In the previous section we described the t-test as the standard method to test a hypothesis on a single coefficient in a regression. In many contexts, however, we want to simultaneously assess a set of coefficients. In the normal regression model, this can be done by an F test which can be derived from the likelihood ratio test.

Partition the regressors as $X = (X_1', X_2')$ and similarly partition the coefficient vector as $\beta = (\beta_1', \beta_2')'$. The regression model can be written as

$$Y = X_1'\beta_1 + X_2'\beta_2 + e. \tag{5.15}$$

Let $k = \dim(X)$, $k_1 = \dim(X_1)$, and $q = \dim(X_2)$, so that $k = k_1 + q$. Partition the variables so that the hypothesis is that the second set of coefficients are zero, or

$$\mathbb{H}_0 : \beta_2 = 0. \tag{5.16}$$

If $\mathbb{H}_0$ is true then the regressors X_2 can be omitted from the regression. In this case we can write (5.15) as

$$Y = X_1'\beta_1 + e. \tag{5.17}$$

We call (5.17) the null model. The alternative hypothesis is that at least one element of β_2 is non-zero and is written as $\mathbb{H}_1 : \beta_2 \neq 0$.

When models are estimated by maximum likelihood a well-accepted testing procedure is to reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ for large values of the Likelihood Ratio – the ratio of the maximized likelihood function under $\mathbb{H}_1$ and $\mathbb{H}_0$, respectively. We now construct this statistic in the normal regression model. Recall from (5.6) that the maximized log-likelihood equals

$$\ell_n\left(\widehat{\beta}, \widehat{\sigma}^2\right) = -\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2\right) - \frac{n}{2}.$$

We similarly calculate the maximized log-likelihood for the constrained model (5.17). By the same steps for derivation of the unconstrained MLE we find that the MLE for (5.17) is OLS of Y on X_1. We can write this estimator as

$$\widetilde{\beta}_1 = \left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'\boldsymbol{Y}$$

with residual $\widetilde{e}_i = Y_i - X_{1i}'\widetilde{\beta}_1$ and error variance estimate $\widetilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widetilde{e}_i^2$. We use tildes "~" rather than hats "^" above the constrained estimates to distinguish them from the unconstrained estimates. You can calculate similar to (5.6) that the maximized constrained log-likelihood is

$$\ell_n\left(\widetilde{\beta}_1, \widetilde{\sigma}^2\right) = -\frac{n}{2}\log\left(2\pi\widetilde{\sigma}^2\right) - \frac{n}{2}.$$

A classic testing procedure is to reject $\mathbb{H}_0$ for large values of the ratio of the maximized likelihoods. Equivalently the test rejects $\mathbb{H}_0$ for large values of twice the difference in the log-likelihood functions. (Multiplying the likelihood difference by two turns out to be a useful scaling.) This equals

$$\begin{aligned}
\mathrm{LR} &= 2\left(\ell_n\left(\widehat{\beta}, \widehat{\sigma}^2\right) - \ell_n\left(\widetilde{\beta}_1, \widetilde{\sigma}^2\right)\right) \\
&= 2\left(\left(-\frac{n}{2}\log\left(2\pi\widehat{\sigma}^2\right) - \frac{n}{2}\right) - \left(-\frac{n}{2}\log\left(2\pi\widetilde{\sigma}^2\right) - \frac{n}{2}\right)\right) \\
&= n\log\left(\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2}\right). \qquad (5.18)
\end{aligned}$$

The likelihood ratio test rejects $\mathbb{H}_0$ for large values of LR, or equivalently (see Exercise 5.10) for large values of

$$\mathrm{F} = \frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)/q}{\widehat{\sigma}^2/(n-k)}. \qquad (5.19)$$

This is known as the F statistic for the test of hypothesis $\mathbb{H}_0$ against $\mathbb{H}_1$.

To develop an appropriate critical value we need the null distribution of F. Recall from (3.28) that $n\widehat{\sigma}^2 = \boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P}$ with $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$. Similarly, under $\mathbb{H}_0$, $n\widetilde{\sigma}^2 = \boldsymbol{e}'\boldsymbol{M}_1\boldsymbol{e}$ where $\boldsymbol{M} = \boldsymbol{I}_n - \boldsymbol{P}_1$ with $\boldsymbol{P}_1 = \boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{X}_1\right)^{-1}\boldsymbol{X}_1'$. You can calculate that $\boldsymbol{M}_1 - \boldsymbol{M} = \boldsymbol{P} - \boldsymbol{P}_1$ is idempotent with rank q. Furthermore, $(\boldsymbol{M}_1 - \boldsymbol{M})\boldsymbol{M} = 0$. It follows that $\boldsymbol{e}'(\boldsymbol{M}_1 - \boldsymbol{M})\boldsymbol{e} \sim \chi_q^2$ and is independent of $\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}$. Hence

$$\mathrm{F} = \frac{\boldsymbol{e}'(\boldsymbol{M}_1 - \boldsymbol{M})\boldsymbol{e}/q}{\boldsymbol{e}'\boldsymbol{M}\boldsymbol{e}/(n-k)} \sim \frac{\chi_q^2/q}{\chi_{n-k}^2/(n-k)} \sim F_{q,n-k}$$

an exact F distribution with degrees of freedom q and $n-k$, respectively. Thus under $\mathbb{H}_0$, the F statistic has an exact F distribution.

The critical values are selected from the upper tail of the F distribution. For a given significance level α (typically $\alpha = 0.05$) we select the critical value c so that $\mathbb{P}\left[F_{q,n-k} \geq c\right] = \alpha$. For example, in MATLAB the expression is `finv(1-α,q,n-k)`. The test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $F > c$ and does not reject $\mathbb{H}_0$ otherwise. The p-value of the test is $p = 1 - G_{q,n-k}(F)$ where $G_{q,n-k}(u)$ is the $F_{q,n-k}$ distribution function. In MATLAB, the p-value is computed as `1-fcdf(f,q,n-k)`. It is equivalent to reject $\mathbb{H}_0$ if $F > c$ or $p < \alpha$.

In Stata, the command to test multiple coefficients takes the form '`test X1 X2`' where X1 and X2 are the names of the variables whose coefficients are tested. Stata then reports the F statistic for the hypothesis that the coefficients are jointly zero along with the p-value calculated using the F distribution.

Theorem 5.13 In the normal regression model, if the null hypothesis (5.16) is true, then for F defined in (5.19), $F \sim F_{q,n-k}$. If c is set so that $\mathbb{P}\left[F_{q,n-k} \geq c\right] = \alpha$ then the test "Reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $F > c$" has significance level α.

Theorem 5.13 justifies the F test in the normal regression model with critical values from the F distribution.

5.14 Information Bound for Normal Regression

This section requires a familiarity with the theory of the Cramér-Rao Lower Bound. See Chapter 10 of *Probability and Statistics for Economists*.

The likelihood scores for the normal regression model are

$$\frac{\partial}{\partial \beta}\ell_n(\beta,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^n X_i\left(Y_i - X_i'\beta\right) = \frac{1}{\sigma^2}\sum_{i=1}^n X_i e_i$$

and

$$\frac{\partial}{\partial \sigma^2}\ell_n(\beta,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^n \left(Y_i - X_i'\beta\right)^2 = \frac{1}{2\sigma^4}\sum_{i=1}^n \left(e_i^2 - \sigma^2\right).$$

It follows that the information matrix is

$$\mathscr{I} = \operatorname{var}\left[\begin{array}{c} \frac{\partial}{\partial\beta}\ell(\beta,\sigma^2) \\ \frac{\partial}{\partial\sigma^2}\ell(\beta,\sigma^2) \end{array}\middle| \boldsymbol{X}\right] = \begin{pmatrix} \frac{1}{\sigma^2}\boldsymbol{X}'\boldsymbol{X} & 0 \\ 0 & \dfrac{2\sigma^4}{n} \end{pmatrix} \tag{5.20}$$

(see Exercise 5.11). The Cramér-Rao Lower Bound is

$$\mathscr{I}^{-1} = \begin{pmatrix} \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} & 0 \\ 0 & \dfrac{2\sigma^4}{n} \end{pmatrix}.$$

This shows that the lower bound for estimation of β is $\sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ and the lower bound for σ^2 is $2\sigma^4/n$.

The unbiased variance estimator s^2 of σ^2 has variance $2\sigma^4/(n-k)$ (see Exercise 5.12) which is larger than the Cramér-Rao lower bound $2\sigma^4/n$. Thus in contrast to the coefficient estimator, the variance estimator is not Cramér-Rao efficient.

5.15 Exercises

Exercise 5.1 Show that if $Q \sim \chi^2_r$, then $\mathbb{E}[Q] = r$ and $\operatorname{var}[Q] = 2r$.

Hint: Use the representation $Q = \sum_{i=1}^n Z_i^2$ with Z_i independent $\mathrm{N}(0,1)$.

Exercise 5.2 Show that if $\boldsymbol{e} \sim \mathrm{N}\left(0, \boldsymbol{I}_n\sigma^2\right)$ and $\boldsymbol{H}'\boldsymbol{H} = \boldsymbol{I}_n$ then $\boldsymbol{u} = \boldsymbol{H}'\boldsymbol{e} \sim \mathrm{N}\left(0, \boldsymbol{I}_n\sigma^2\right)$.

Exercise 5.3 Show that if $\boldsymbol{e} \sim \mathrm{N}(0,\Sigma)$ and $\Sigma = \boldsymbol{A}\boldsymbol{A}'$ then $\boldsymbol{u} = \boldsymbol{A}^{-1}\boldsymbol{e} \sim \mathrm{N}(0, \boldsymbol{I}_n)$.

Exercise 5.4 Show that $\operatorname{argmax}_{\theta\in\Theta}\ell_n(\theta) = \operatorname{argmax}_{\theta\in\Theta} L_n(\theta)$.

Exercise 5.5 For the regression in-sample predicted values $\widehat{Y}_i$ show that $\widehat{Y}_i \mid \boldsymbol{X} \sim \mathrm{N}\left(X_i'\beta, \sigma^2 h_{ii}\right)$ where h_{ii} are the leverage values (3.40).

Exercise 5.6 In the normal regression model show that the leave-one out prediction errors $\widetilde{e}_i$ and the standardized residuals $\overline{e}_i$ are independent of $\widehat{\beta}$, conditional on $\boldsymbol{X}$.

Hint: Use (3.45) and (4.29).

Exercise 5.7 In the normal regression model show that the robust covariance matrices $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC0}}$, $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC1}}$, $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC2}}$, and $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC3}}$ are independent of the OLS estimator $\widehat{\beta}$, conditional on $\boldsymbol{X}$.

Exercise 5.8 Let $F(u)$ be the distribution function of a random variable X whose density is symmetric about zero. (This includes the standard normal and the student t.) Show that $F(-u) = 1 - F(u)$.

Exercise 5.9 Let $\widehat{C}_\beta = [L, U]$ be a $1-\alpha$ confidence interval for β, and consider the transformation $\theta = g(\beta)$ where $g(\cdot)$ is monotonically increasing. Consider the confidence interval $\widehat{C}_\theta = [g(L), g(U)]$ for θ. Show that $\mathbb{P}\left[\theta \in \widehat{C}_\theta\right] = \mathbb{P}\left[\beta \in \widehat{C}_\beta\right]$. Use this result to develop a confidence interval for σ.

Exercise 5.10 Show that the test "Reject $\mathbb{H}_0$ if LR $\geq c_1$" for LR defined in (5.18), and the test "Reject $\mathbb{H}_0$ if F $\geq c_2$" for F defined in (5.19), yield the same decisions if $c_2 = \left(\exp(c_1/n) - 1\right)(n-k)/q$. Does this mean that the two tests are equivalent?

Exercise 5.11 Show (5.20).

Exercise 5.12 In the normal regression model let s^2 be the unbiased estimator of the error variance σ^2 from (4.31).

(a) Show that $\operatorname{var}\left[s^2\right] = 2\sigma^4/(n-k)$.

(b) Show that $\operatorname{var}\left[s^2\right]$ is strictly larger than the Cramér-Rao Lower Bound for σ^2.

Part II

Large Sample Methods

Chapter 6

A Review of Large Sample Asymptotics

6.1 Introduction

The most widely-used tool in sampling theory is large sample asymptotics. By "asymptotics" we mean approximating a finite-sample sampling distribution by taking its limit as the sample size diverges to infinity. In this chapter we provide a brief review of the main results of large sample asymptotics. It is meant as a reference, not as a teaching guide. Asymptotic theory is covered in detail in Chapters 7-9 of *Probability and Statistics for Economists*. If you have not previous studied asymptotic theory in detail you should study these chapters before proceeding.

6.2 Modes of Convergence

Definition 6.1 A sequence of random vectors $Z_n \in \mathbb{R}^k$ **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 $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left[\|Z_n - Z\| \leq \delta\right] = 1. \tag{6.1}$$

We call Z the **probability limit** (or **plim**) of Z_n.

The above definition treats random variables and random vectors simultaneously using the vector norm. It is useful to know that for a random vector, (6.1) holds if and only if each element in the vector converges in probability to its limit.

Definition 6.2 Let Z_n be a sequence of random vectors with distributions $F_n(u) = \mathbb{P}\left[Z_n \leq u\right]$. We say that Z_n **converges in distribution** to Z as $n \to \infty$, denoted $Z_n \xrightarrow[d]{} Z$, if for all u at which $F(u) = \mathbb{P}\left[Z \leq u\right]$ is continuous, $F_n(u) \to F(u)$ as $n \to \infty$. We refer to Z and its distribution $F(u)$ as the **asymptotic distribution**, **large sample distribution**, or **limit distribution** of Z_n.

6.3 Weak Law of Large Numbers

Theorem 6.1 Weak Law of Large Numbers (WLLN)
If $Y_i \in \mathbb{R}^k$ are i.i.d. 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 WLLN shows that the sample mean $\overline{Y}$ converges in probability to the true population expectation μ. The result applies to any transformation of a random vector with a finite mean.

Theorem 6.2 If $Y_i \in \mathbb{R}^k$ are i.i.d., $h(y) : \mathbb{R}^k \to \mathbb{R}^q$, and $\mathbb{E}\|h(Y)\| < \infty$, then $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(Y_i) \xrightarrow[p]{} \mu = \mathbb{E}[h(Y)]$ as $n \to \infty$.

An estimator which converges in probability to the population value is called **consistent**.

Definition 6.3 An estimator $\widehat{\theta}$ of θ is **consistent** if $\widehat{\theta} \xrightarrow[p]{} \theta$ as $n \to \infty$.

6.4 Central Limit Theorem

Theorem 6.3 Multivariate Lindeberg-Lévy Central Limit Theorem (CLT). If $Y_i \in \mathbb{R}^k$ are i.i.d. and $\mathbb{E}\|Y\|^2 < \infty$, then as $n \to \infty$

$$\sqrt{n}\left(\overline{Y} - \mu\right) \xrightarrow[d]{} \mathrm{N}(0, \boldsymbol{V})$$

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

The central limit theorem shows that the distribution of the sample mean is approximately normal in large samples. For some applications it may be useful to notice that Theorem 6.3 does not impose any restrictions on $\boldsymbol{V}$ other than that the elements are finite. Therefore this result allows for the possibility of singular $\boldsymbol{V}$.

The following two generalizations allow for heterogeneous random variables.

Theorem 6.4 Multivariate Lindeberg CLT. Suppose that for all n, $Y_{ni} \in \mathbb{R}^k$, $i = 1, ..., r_n$, are independent but not necessarily identically distributed with expectations $\mathbb{E}[Y_{ni}] = 0$ and variance matrices $\boldsymbol{V}_{ni} = \mathbb{E}\left[Y_{ni} Y_{ni}'\right]$. Set $\overline{\boldsymbol{V}}_n = \sum_{i=1}^n \boldsymbol{V}_{ni}$. Suppose $\nu_n^2 = \lambda_{\min}(\overline{\boldsymbol{V}}_n) > 0$ and for all $\epsilon > 0$

$$\lim_{n\to\infty} \frac{1}{\nu_n^2} \sum_{i=1}^{r_n} \mathbb{E}\left[\|Y_{ni}\|^2 \mathbb{1}\left\{\|Y_{ni}\|^2 \geq \epsilon \nu_n^2\right\}\right] = 0. \tag{6.2}$$

Then as $n \to \infty$

$$\overline{\boldsymbol{V}}_n^{-1/2} \sum_{i=1}^{r_n} Y_{ni} \xrightarrow[d]{} \mathrm{N}(0, \boldsymbol{I}_k).$$

Theorem 6.5 Suppose $Y_{ni} \in \mathbb{R}^k$ are independent but not necessarily identically distributed with expectations $\mathbb{E}[Y_{ni}] = 0$ and variance matrices $\boldsymbol{V}_{ni} = \mathbb{E}\left[Y_{ni} Y_{ni}'\right]$. Suppose

$$\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{V}_{ni} \to \boldsymbol{V} > 0$$

and for some $\delta > 0$

$$\sup_{n,i} \mathbb{E}\|Y_{ni}\|^{2+\delta} < \infty. \tag{6.3}$$

Then as $n \to \infty$

$$\sqrt{n}\,\overline{Y} \xrightarrow[d]{} \mathrm{N}(0, \boldsymbol{V}).$$

6.5 Continuous Mapping Theorem and Delta Method

Continuous functions are limit-preserving. There are two forms of the continuous mapping theorem, for convergence in probability and convergence in distribution.

Theorem 6.6 Continuous Mapping Theorem (CMT). Let $Z_n \in \mathbb{R}^k$ and $g(u) : \mathbb{R}^k \to \mathbb{R}^q$. If $Z_n \xrightarrow[p]{} c$ as $n \to \infty$ and $g(u)$ is continuous at c then $g(Z_n) \xrightarrow[p]{} g(c)$ as $n \to \infty$.

Theorem 6.7 Continuous Mapping Theorem. If $Z_n \xrightarrow[d]{} Z$ as $n \to \infty$ and $g : \mathbb{R}^m \to \mathbb{R}^k$ has the set of discontinuity points D_g such that $\mathbb{P}\left[Z \in D_g\right] = 0$, then $g(Z_n) \xrightarrow[d]{} g(Z)$ as $n \to \infty$.

Differentiable functions of asymptotically normal random estimators are asymptotically normal.

Theorem 6.8 Delta Method. Let $\mu \in \mathbb{R}^k$ and $g(u) : \mathbb{R}^k \to \mathbb{R}^q$. If $\sqrt{n}\left(\widehat{\mu} - \mu\right) \underset{d}{\longrightarrow} \xi$, where $g(u)$ is continuously differentiable in a neighborhood of μ, then as $n \to \infty$

$$\sqrt{n}\left(g\left(\widehat{\mu}\right) - g(\mu)\right) \underset{d}{\longrightarrow} \boldsymbol{G}'\xi \tag{6.4}$$

where $\boldsymbol{G}(u) = \frac{\partial}{\partial u} g(u)'$ and $\boldsymbol{G} = \boldsymbol{G}(\mu)$. In particular, if $\xi \sim \mathrm{N}(0, \boldsymbol{V})$ then as $n \to \infty$

$$\sqrt{n}\left(g\left(\widehat{\mu}\right) - g(\mu)\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}\right). \tag{6.5}$$

6.6 Smooth Function Model

The smooth function model is $\theta = g\left(\mu\right)$ where $\mu = \mathbb{E}\left[h\left(Y\right)\right]$ and $g\left(\mu\right)$ is smooth in a suitable sense.

The parameter $\theta = g\left(\mu\right)$ is not a population moment so it does not have a direct moment estimator. Instead, it is common to use a **plug-in estimator** formed by replacing the unknown μ with its point estimator $\widehat{\mu}$ and then "plugging" this into the expression for θ. The first step is the sample mean $\widehat{\mu} = n^{-1}\sum_{i=1}^{n} h\left(Y_i\right)$. The second step is the transformation $\widehat{\theta} = g\left(\widehat{\mu}\right)$. The hat "^" indicates that $\widehat{\theta}$ is a sample estimator of θ. The smooth function model includes a broad class of estimators including sample variances and the least squares estimator.

Theorem 6.9 If $Y_i \in \mathbb{R}^m$ are i.i.d., $h(u) : \mathbb{R}^m \to \mathbb{R}^k$, $\mathbb{E}\left\|h\left(Y\right)\right\| < \infty$, and $g\left(u\right) : \mathbb{R}^k \to \mathbb{R}^q$ is continuous at μ, then $\widehat{\theta} \underset{p}{\longrightarrow} \theta$ as $n \to \infty$.

Theorem 6.10 If $Y_i \in \mathbb{R}^m$ are i.i.d., $h(u) : \mathbb{R}^m \to \mathbb{R}^k$, $\mathbb{E}\left\|h\left(Y\right)\right\|^2 < \infty$, $g\left(u\right) : \mathbb{R}^k \to \mathbb{R}^q$, and $\boldsymbol{G}\left(u\right) = \dfrac{\partial}{\partial u} g(u)'$ is continuous in a neighborhood of μ, then as $n \to \infty$

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\theta\right)$$

where $\boldsymbol{V}_\theta = \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}$, $\boldsymbol{V} = \mathbb{E}\left[\left(h\left(Y\right) - \mu\right)\left(h\left(Y\right) - \mu\right)'\right]$, and $\boldsymbol{G} = \boldsymbol{G}\left(\mu\right)$.

Theorem 6.9 establishes the consistency of $\widehat{\theta}$ for θ and Theorem 6.10 establishes its asymptotic normality. It is instructive to compare the conditions. Consistency requires that $h\left(Y\right)$ has a finite expectation; asymptotic normality requires that $h\left(Y\right)$ has a finite variance. Consistency requires that $g(u)$ be continuous; asymptotic normality requires that $g(u)$ is continuously differentiable.

6.7 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 common notation.

Let Z_n and a_n, $n = 1, 2, ...$ be sequences of random variables and constants. The notation

$$Z_n = o_p(1)$$

("small oh-P-one") means that $Z_n \underset{p}{\longrightarrow} 0$ as $n \to \infty$. We also write

$$Z_n = o_p(a_n)$$

if $a_n^{-1} Z_n = o_p(1)$.

Similarly, the notation $Z_n = O_p(1)$ ("big oh-P-one") means that Z_n is bounded in probability. Precisely, for any $\epsilon > 0$ there is a constant $M_\epsilon < \infty$ such that

$$\limsup_{n\to\infty} \mathbb{P}[|Z_n| > M_\epsilon] \le \epsilon.$$

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 $a_n / b_n \to 0$.

A random sequence with a bounded moment is stochastically bounded.

Theorem 6.11 If Z_n is a random vector which satisfies $\mathbb{E}\|Z_n\|^\delta = O(a_n)$ for some sequence a_n and $\delta > 0$, then $Z_n = O_p(a_n^{1/\delta})$. Similarly, $\mathbb{E}\|Z_n\|^\delta = o(a_n)$ implies $Z_n = o_p(a_n^{1/\delta})$.

There are many simple rules for manipulating $o_p(1)$ and $O_p(1)$ sequences which can be deduced from the continuous mapping theorem. For example,

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

6.8 Convergence of Moments

We give a sufficient condition for the existence of the mean of the asymptotic distribution, define uniform integrability, provide a primitive condition for uniform integrability, and show that uniform integrability is the key condition under which $\mathbb{E}[Z_n]$ converges to $\mathbb{E}[Z]$.

Theorem 6.12 If $Z_n \xrightarrow[d]{} Z$ and $\mathbb{E}\|Z_n\| \leq C$ then $\mathbb{E}\|Z\| \leq C$.

Definition 6.4 The random vector Z_n is **uniformly integrable** as $n \to \infty$ if

$$\lim_{M\to\infty} \limsup_{n\to\infty} \mathbb{E}[\|Z_n\| \mathbb{1}\{\|Z_n\| > M\}] = 0.$$

Theorem 6.13 If for some $\delta > 0$, $\mathbb{E}\|Z_n\|^{1+\delta} \leq C < \infty$, then Z_n is uniformly integrable.

Theorem 6.14 If $Z_n \xrightarrow[d]{} Z$ and Z_n is uniformly integrable then $\mathbb{E}[Z_n] \longrightarrow \mathbb{E}[Z]$.

The following is a uniform stochastic bound.

Theorem 6.15 If $|Y_i|^r$ is uniformly integrable, then as $n \to \infty$

$$n^{-1/r} \max_{1\leq i\leq n} |Y_i| \xrightarrow[p]{} 0. \tag{6.6}$$

Equation (6.6) implies that if Y has r finite moments then the largest observation will diverge at a rate slower than $n^{1/r}$. The higher the moments, the slower the rate of divergence.

Chapter 7

Asymptotic Theory for Least Squares

7.1 Introduction

It turns out that the asymptotic theory of least squares estimation applies equally to the projection model and the linear CEF model. Therefore the results in this chapter will be stated for the broader projection model described in Section 2.18. Recall that the model is $Y = X'\beta + e$ with the linear projection coefficient $\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right]$.

Maintained assumptions in this chapter will be random sampling (Assumption 1.2) and finite second moments (Assumption 2.1). We restate these here for clarity.

Assumption 7.1

1. The variables (Y_i, X_i), $i = 1, ..., n$, are i.i.d.
2. $\mathbb{E}\left[Y^2\right] < \infty$.
3. $\mathbb{E}\|X\|^2 < \infty$.
4. $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$ is positive definite.

The distributional results will require a strengthening of these assumptions to finite fourth moments. We discuss the specific conditions in Section 7.3.

7.2 Consistency of Least Squares Estimator

In this section we use the weak law of large numbers (WLLN, Theorem 6.1 and Theorem 6.2) and continuous mapping theorem (CMT, Theorem 6.6) to show that the least squares estimator $\widehat{\beta}$ is consistent for the projection coefficient β.

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

$$\widehat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{XY}$$

is a function of the sample moments $\widehat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ and $\widehat{\boldsymbol{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$.

Second, by an application of the WLLN these sample moments converge in probability to their population expectations. Specifically, the fact that (Y_i, X_i) are mutually i.i.d. implies that any function of (Y_i, X_i) is i.i.d., including $X_i X_i'$ and $X_i Y_i$. These variables also have finite expectations under Assumption 7.1. Under these conditions, the WLLN implies that as $n \to \infty$,

$$\widehat{\boldsymbol{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \underset{p}{\longrightarrow} \mathbb{E}\left[XX'\right] = \boldsymbol{Q}_{XX} \tag{7.1}$$

and

$$\widehat{\boldsymbol{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \underset{p}{\longrightarrow} \mathbb{E}\left[XY\right] = \boldsymbol{Q}_{XY}.$$

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

$$\widehat{\beta} = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{XY} \underset{p}{\longrightarrow} \boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} = \beta. \tag{7.2}$$

We have shown that $\widehat{\beta} \underset{p}{\longrightarrow} \beta$ as $n \to \infty$. In words, the OLS estimator converges in probability to the projection coefficient vector β as the sample size n gets large.

To fully understand the application of the CMT we walk through it in detail. We can write

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

where $g(\boldsymbol{A}, \boldsymbol{b}) = \boldsymbol{A}^{-1}\boldsymbol{b}$ is a function of $\boldsymbol{A}$ and $\boldsymbol{b}$. The function $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b})$ is a continuous function of $\boldsymbol{A}$ and $\boldsymbol{b}$ at all values of the arguments such that $\boldsymbol{A}^{-1}$ exists. Assumption 7.1 specifies that $\boldsymbol{Q}_{XX}$ is positive definite, which means that $\boldsymbol{Q}_{XX}^{-1}$ exists. Thus $\boldsymbol{g}(\boldsymbol{A}, \boldsymbol{b})$ is continuous at $\boldsymbol{A} = \boldsymbol{Q}_{XX}$. This justifies the application of the CMT in (7.2).

For a slightly different demonstration of (7.2) recall that (4.6) implies that

$$\widehat{\beta} - \beta = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{Xe} \tag{7.3}$$

where

$$\widehat{\boldsymbol{Q}}_{Xe} = \frac{1}{n}\sum_{i=1}^{n} X_i e_i.$$

The WLLN and (2.25) imply

$$\widehat{\boldsymbol{Q}}_{Xe} \underset{p}{\longrightarrow} \mathbb{E}\left[Xe\right] = 0.$$

Therefore

$$\widehat{\beta} - \beta = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{Q}}_{Xe} \underset{p}{\longrightarrow} \boldsymbol{Q}_{XX}^{-1}0 = 0$$

which is the same as $\widehat{\beta} \underset{p}{\longrightarrow} \beta$.

Theorem 7.1 Consistency of Least Squares. Under Assumption 7.1, $\widehat{\boldsymbol{Q}}_{XX} \xrightarrow[p]{} \boldsymbol{Q}_{XX}$, $\widehat{\boldsymbol{Q}}_{XY} \xrightarrow[p]{} \boldsymbol{Q}_{XY}$, $\widehat{\boldsymbol{Q}}_{XX}^{-1} \xrightarrow[p]{} \boldsymbol{Q}_{XX}^{-1}$, $\widehat{\boldsymbol{Q}}_{Xe} \xrightarrow[p]{} 0$, and $\widehat{\beta} \xrightarrow[p]{} \beta$ as $n \to \infty$.

Theorem 7.1 states that the OLS estimator $\widehat{\beta}$ converges in probability to β as n increases and thus $\widehat{\beta}$ is consistent for β. In the stochastic order notation, Theorem 7.1 can be equivalently written as

$$\widehat{\beta} = \beta + o_p(1). \tag{7.4}$$

To illustrate the effect of sample size on the least squares estimator consider the least squares regression

$$\log(wage) = \beta_1 education + \beta_2 experience + \beta_3 experience^2 + \beta_4 + e.$$

We use the sample of 24,344 white men from the March 2009 CPS. We randomly sorted the observations and sequentially estimated the model by least squares starting with the first 5 observations and continuing until the full sample is used. The sequence of estimates are displayed in Figure 7.1. You can see how the least squares estimate changes with the sample size. As the number of observations increases it settles down to the full-sample estimate $\widehat{\beta}_1 = 0.114$.

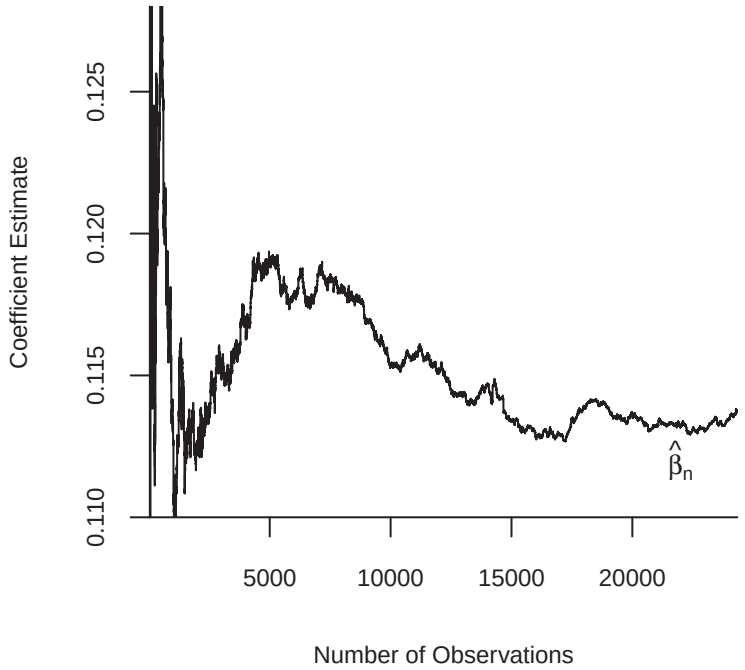


Figure 7.1: The Least-Squares Estimator as a Function of Sample Size

7.3 Asymptotic Normality

We started this chapter discussing the need for an approximation to the distribution of the OLS estimator $\widehat{\beta}$. In Section 7.2 we showed that $\widehat{\beta}$ converges in probability to β. Consistency is a good first step, but in itself does not describe the distribution of the estimator. In this section we derive an approximation typically called the **asymptotic distribution**.

The derivation starts by writing the estimator as a function of sample moments. One of the moments must be written as a sum of zero-mean random vectors and normalized so that the central limit theorem can be applied. The steps are as follows.

Take equation (7.3) and multiply it by $\sqrt{n}$. This yields the expression

$$\sqrt{n}\left(\widehat{\beta}-\beta\right)=\left(\frac{1}{n}\sum_{i=1}^{n}X_iX_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_ie_i\right). \tag{7.5}$$

This shows that the normalized and centered estimator $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ is a function of the sample average $n^{-1}\sum_{i=1}^{n}X_iX_i'$ and the normalized sample average $n^{-1/2}\sum_{i=1}^{n}X_ie_i$.

The random pairs (Y_i,X_i) are i.i.d., meaning that they are independent across i and identically distributed. Any function of (Y_i,X_i) is also i.i.d. This includes $e_i=Y_i-X_i'\beta$ and the product X_ie_i. The latter is mean-zero ($\mathbb{E}[Xe]=0$) and has $k\times k$ covariance matrix

$$\Omega=\mathbb{E}\left[(Xe)(Xe)'\right]=\mathbb{E}\left[XX'e^2\right].$$

We show below that Ω has finite elements under a strengthening of Assumption 7.1. Since X_ie_i is i.i.d., mean zero, and finite variance, the central limit theorem (Theorem 6.3) implies

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}X_ie_i\underset{d}{\longrightarrow}\mathrm{N}(0,\Omega).$$

We state the required conditions here.

Assumption 7.2

1. The variables (Y_i,X_i), $i=1,...,n$, are i.i.d..
2. $\mathbb{E}\left[Y^4\right]<\infty$.
3. $\mathbb{E}\|X\|^4<\infty$.
4. $\boldsymbol{Q}_{XX}=\mathbb{E}\left[XX'\right]$ is positive definite.

Assumption 7.2 implies that $\Omega<\infty$. To see this, take its $j\ell^{th}$ element, $\mathbb{E}\left[X_jX_\ell e^2\right]$. Theorem 2.9.6 shows that $\mathbb{E}\left[e^4\right]<\infty$. By the expectation inequality (B.30), the $j\ell^{th}$ element of Ω is bounded by

$$\left|\mathbb{E}\left[X_jX_\ell e^2\right]\right|\leq\mathbb{E}\left|X_jX_\ell e^2\right|=\mathbb{E}\left[\left|X_j\right|\left|X_\ell\right|e^2\right].$$

By two applications of the Cauchy-Schwarz inequality (B.32), this is smaller than

$$\left(\mathbb{E}\left[X_j^2X_\ell^2\right]\right)^{1/2}\left(\mathbb{E}\left[e^4\right]\right)^{1/2}\leq\left(\mathbb{E}\left[X_j^4\right]\right)^{1/4}\left(\mathbb{E}\left[X_\ell^4\right]\right)^{1/4}\left(\mathbb{E}\left[e^4\right]\right)^{1/2}<\infty$$

where the finiteness holds under Assumption 7.2.2 and 7.2.3. Thus $\Omega < \infty$.

An alternative way to show that the elements of Ω are finite is by using a matrix norm $\|\cdot\|$ (See Appendix A.23). Then by the expectation inequality, the Cauchy-Schwarz inequality, Assumption 7.2.3, and $\mathbb{E}\left[e^4\right] < \infty$,

$$\|\Omega\| \leq \mathbb{E}\left\|XX'e^2\right\| = \mathbb{E}\left[\|X\|^2 e^2\right] \leq \left(\mathbb{E}\|X\|^4\right)^{1/2}\left(\mathbb{E}\left[e^4\right]\right)^{1/2} < \infty.$$

This is a more compact argument (often described as more *elegant*) but such manipulations should not be done without understanding the notation and the applicability of each step of the argument.

Regardless, the finiteness of the covariance matrix means that we can apply the multivariate CLT (Theorem 6.3).

Theorem 7.2 Assumption 7.2 implies that

$$\Omega < \infty \tag{7.6}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i \underset{d}{\longrightarrow} \mathrm{N}(0, \Omega) \tag{7.7}$$

as $n \to \infty$.

Putting together (7.1), (7.5), and (7.7),

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \underset{d}{\longrightarrow} \boldsymbol{Q}_{XX}^{-1}\mathrm{N}(0, \Omega) = \mathrm{N}\left(0, \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}\right)$$

as $n \to \infty$. The final equality follows from the property that linear combinations of normal vectors are also normal (Theorem 5.2).

We have derived the asymptotic normal approximation to the distribution of the least squares estimator.

Theorem 7.3 Asymptotic Normality of Least Squares Estimator
Under Assumption 7.2, as $n \to \infty$

$$\sqrt{n}\left(\widehat{\beta} - \beta\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right)$$

where $\boldsymbol{Q}_{XX} = \mathbb{E}\left[XX'\right]$, $\Omega = \mathbb{E}\left[XX'e^2\right]$, and

$$\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}. \tag{7.8}$$

In the stochastic order notation, Theorem 7.3 implies that $\widehat{\beta} = \beta + O_p(n^{-1/2})$ which is stronger than (7.4).

The matrix $\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$ is the variance of the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$. Consequently, $\boldsymbol{V}_\beta$ is often referred to as the **asymptotic covariance matrix** of $\widehat{\beta}$. The expression $\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$ is called a **sandwich** form as the matrix Ω is sandwiched between two copies of $\boldsymbol{Q}_{XX}^{-1}$.

It is useful to compare the variance of the asymptotic distribution given in (7.8) and the finite-sample conditional variance in the CEF model as given in (4.10):

$$\boldsymbol{V}_{\widehat{\beta}} = \operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}. \tag{7.9}$$

Notice that $\boldsymbol{V}_{\widehat{\beta}}$ is the exact conditional variance of $\widehat{\beta}$ and $\boldsymbol{V}_{\beta}$ is the asymptotic variance of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$. Thus $\boldsymbol{V}_{\beta}$ should be (roughly) n times as large as $\boldsymbol{V}_{\widehat{\beta}}$, or $\boldsymbol{V}_{\beta} \approx n\boldsymbol{V}_{\widehat{\beta}}$. Indeed, multiplying (7.9) by n and distributing we find

$$n\boldsymbol{V}_{\widehat{\beta}} = \left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$

which looks like an estimator of $\boldsymbol{V}_{\beta}$. Indeed, as $n \to \infty$, $n\boldsymbol{V}_{\widehat{\beta}} \xrightarrow[p]{} \boldsymbol{V}_{\beta}$. The expression $\boldsymbol{V}_{\widehat{\beta}}$ is useful for practical inference (such as computation of standard errors and tests) as it is the variance of the estimator $\widehat{\beta}$, while $\boldsymbol{V}_{\beta}$ 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 Ω and $\boldsymbol{V}_{\beta}$ simplify. Suppose that

$$\operatorname{cov}(XX', e^2) = 0. \tag{7.10}$$

Condition (7.10) holds in the homoskedastic linear regression model but is somewhat broader. Under (7.10) the asymptotic variance formulae simplify as

$$\begin{aligned}\Omega &= \mathbb{E}\left[XX'\right]\mathbb{E}\left[e^2\right] = \boldsymbol{Q}_{XX}\sigma^2 \\ \boldsymbol{V}_{\beta} &= \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1} = \boldsymbol{Q}_{XX}^{-1}\sigma^2 \equiv \boldsymbol{V}_{\beta}^{0}.\end{aligned} \tag{7.11}$$

In (7.11) we define $\boldsymbol{V}_{\beta}^{0} = \boldsymbol{Q}_{XX}^{-1}\sigma^2$ whether (7.10) is true or false. When (7.10) is true then $\boldsymbol{V}_{\beta} = \boldsymbol{V}_{\beta}^{0}$, otherwise $\boldsymbol{V}_{\beta} \neq \boldsymbol{V}_{\beta}^{0}$. We call $\boldsymbol{V}_{\beta}^{0}$ the **homoskedastic asymptotic covariance matrix**.

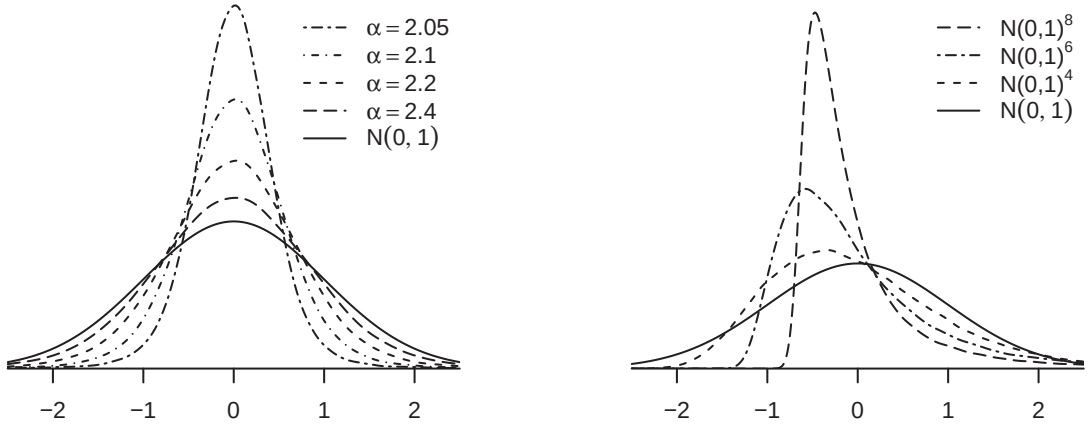
Theorem 7.3 states that the sampling distribution of the least squares estimator, after rescaling, is approximately normal when the sample size n is sufficiently large. This holds true for all joint distributions of (Y, X) which satisfy the conditions of Assumption 7.2. Consequently, asymptotic normality is routinely used to approximate the finite sample distribution of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$.

A difficulty is that for any fixed n the sampling distribution of $\widehat{\beta}$ can be arbitrarily far from the normal distribution. 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 $Y = \beta_1 X + \beta_2 + e$ where X is $\mathrm{N}(0,1)$ and e is independent of X with the Double Pareto density $f(e) = \frac{\alpha}{2}|e|^{-\alpha-1}$, $|e| \geq 1$. If $\alpha > 2$ the error e has zero mean and variance $\alpha/(\alpha-2)$. As α approaches 2, however, its variance diverges to infinity. In this context the normalized least squares slope estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\widehat{\beta}_1-\beta_1\right)$ has the $\mathrm{N}(0,1)$ asymptotic distribution for any $\alpha > 2$. In Figure 7.2(a) we display the finite sample densities of the normalized estimator $\sqrt{n\frac{\alpha-2}{\alpha}}\left(\widehat{\beta}_1-\beta_1\right)$, setting $n = 100$ and varying the parameter α. For $\alpha = 3.0$ the density is very close to the $\mathrm{N}(0,1)$ density. As α diminishes the density changes significantly, concentrating most of the probability mass around zero.

Another example is shown in Figure 7.2(b). Here the model is $Y = \beta + e$ where

$$e = \frac{u^r - \mathbb{E}\left[u^r\right]}{\left(\mathbb{E}\left[u^{2r}\right] - \left(\mathbb{E}\left[u^r\right]\right)^2\right)^{1/2}} \tag{7.12}$$

and $u \sim \mathrm{N}(0,1)$. We show the sampling distribution of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ for $n = 100$, varying $r = 1$, 4, 6 and 8. As r increases, the sampling distribution becomes highly skewed and non-normal. The lesson from Figure 7.2 is that the $\mathrm{N}(0,1)$ asymptotic approximation is never guaranteed to be accurate.



(a) Double Pareto Error

(b) Error Process (7.12)

Figure 7.2: Density of Normalized OLS Estimator

7.4 Joint Distribution

Theorem 7.3 gives the joint asymptotic distribution of the coefficient estimators. We can use the result to study the covariance between the coefficient estimators. For simplicity, take the case of two regressors, no intercept, and homoskedastic error. Assume the regressors are mean zero, variance one, with correlation ρ. Then using the formula for inversion of a 2×2 matrix,

$$\boldsymbol{V}_{\beta}^{0} = \sigma^2 \boldsymbol{Q}_{XX}^{-1} = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}.$$

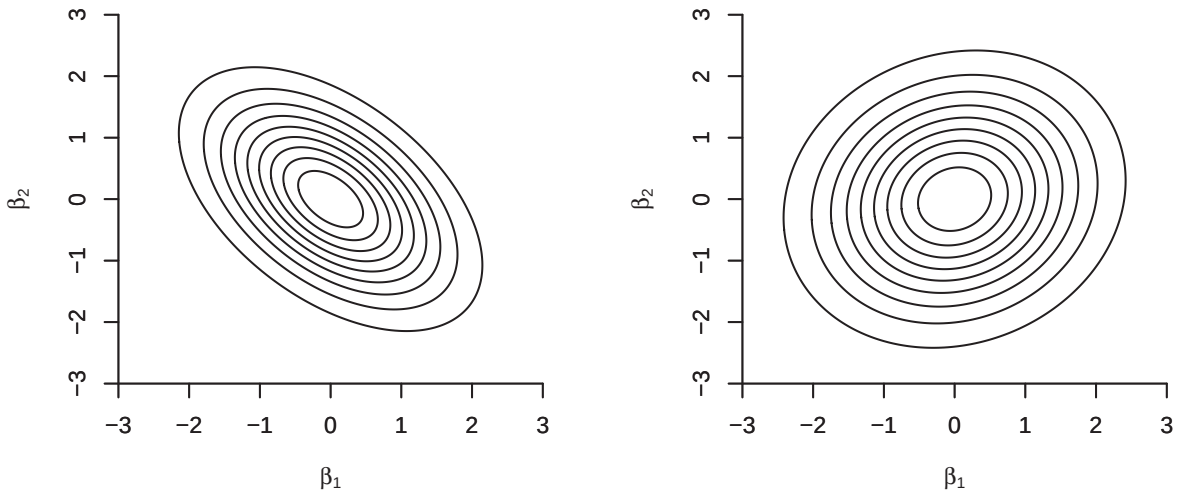
Thus if X_1 and X_2 are positively correlated ($\rho > 0$) then $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are negatively correlated (and vice-versa).

For illustration, Figure 7.3(a) displays the probability contours of the joint asymptotic distribution of $\widehat{\beta}_1 - \beta_1$ and $\widehat{\beta}_2 - \beta_2$ when $\beta_1 = \beta_2 = 0$ and $\rho = 0.5$. The coefficient estimators are negatively correlated because the regressors are positively correlated. This means that if $\widehat{\beta}_1$ is unusually negative, it is likely that $\widehat{\beta}_2$ is unusually positive, or conversely. It is also unlikely that we will observe both $\widehat{\beta}_1$ and $\widehat{\beta}_2$ unusually large and of the same sign.

This finding that the correlation of the regressors is of opposite sign of the correlation of the coefficient estimates is sensitive to the assumption of homoskedasticity. If the errors are heteroskedastic then this relationship is not guaranteed.

This can be seen through a simple constructed example. Suppose that X_1 and X_2 only take the values $\{-1, +1\}$, symmetrically, with $\mathbb{P}[X_1 = X_2 = 1] = \mathbb{P}[X_1 = X_2 = -1] = 3/8$, and $\mathbb{P}[X_1 = 1, X_2 = -1] = \mathbb{P}[X_1 = -1, X_2 = 1] = 1/8$. You can check that the regressors are mean zero, unit variance and correlation 0.5, which is identical with the setting displayed in Figure 7.3(a).

Now suppose that the error is heteroskedastic. Specifically, suppose that $\mathbb{E}\left[e^2 \mid X_1 = X_2\right] = 5/4$ and $\mathbb{E}\left[e^2 \mid X_1 \neq X_2\right] = 1/4$. You can check that $\mathbb{E}\left[e^2\right] = 1$, $\mathbb{E}\left[X_1^2 e^2\right] = \mathbb{E}\left[X_2^2 e^2\right] = 1$ and $\mathbb{E}\left[X_1 X_2 e_i^2\right] = 7/8$. There-



(a) Homoskedastic Case

(b) Heteroskedastic Case

Figure 7.3: Contours of Joint Distribution of $\widehat{\beta}_1$ and $\widehat{\beta}_2$

fore

$$\begin{aligned}\boldsymbol{V}_\beta &= \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1} \\ &= \frac{9}{16}\begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}\begin{bmatrix} 1 & \frac{7}{8} \\ \frac{7}{8} & 1 \end{bmatrix}\begin{bmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \\ &= \frac{4}{3}\begin{bmatrix} 1 & \frac{1}{4} \\ \frac{1}{4} & 1 \end{bmatrix}.\end{aligned}$$

Thus the coefficient estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are positively correlated (their correlation is 1/4.) The joint probability contours of their asymptotic distribution is displayed in Figure 7.3(b). We can see how the two estimators 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 estimators.

We can extend the above analysis to study the covariance between coefficient sub-vectors. For example, partitioning $X' = \left(X_1', X_2'\right)$ and $\beta' = \left(\beta_1', \beta_2'\right)$, we can write the general model as

$$Y = X_1'\beta_1 + X_2'\beta_2 + e$$

and the coefficient estimates as $\widehat{\beta}' = \left(\widehat{\beta}_1', \widehat{\beta}_2'\right)$. Make the partitions

$$\boldsymbol{Q}_{XX} = \begin{bmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{bmatrix}, \qquad \Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}.$$

From (2.43)

$$\boldsymbol{Q}_{XX}^{-1} = \begin{bmatrix} \boldsymbol{Q}_{11\cdot 2}^{-1} & -\boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1} \\ -\boldsymbol{Q}_{22\cdot 1}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1} & \boldsymbol{Q}_{22\cdot 1}^{-1} \end{bmatrix}$$

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

$$\operatorname{cov}\left(\widehat{\beta}_1, \widehat{\beta}_2\right) = -\sigma^2 \boldsymbol{Q}_{11\cdot 2}^{-1}\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}$$

which is a matrix generalization of the two-regressor case.

In general you can show that (Exercise 7.5)

$$\boldsymbol{V}_{\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{bmatrix} \tag{7.13}$$

where

$$\boldsymbol{V}_{11} = \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\Omega_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\Omega_{21} - \Omega_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\Omega_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot 2}^{-1} \tag{7.14}$$

$$\boldsymbol{V}_{21} = \boldsymbol{Q}_{22\cdot 1}^{-1}\left(\Omega_{21} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\Omega_{11} - \Omega_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\Omega_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot 2}^{-1} \tag{7.15}$$

$$\boldsymbol{V}_{22} = \boldsymbol{Q}_{22\cdot 1}^{-1}\left(\Omega_{22} - \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\Omega_{12} - \Omega_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12} + \boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\Omega_{11}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}\right)\boldsymbol{Q}_{22\cdot 1}^{-1}. \tag{7.16}$$

Unfortunately, these expressions are not easily interpretable.

7.5 Consistency of Error Variance Estimators

Using the methods of Section 7.2 we can show that the estimators $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \widehat{e}_i^2$ and $s^2 = (n-k)^{-1}\sum_{i=1}^n \widehat{e}_i^2$ are consistent for σ^2.

The trick is to write the residual $\widehat{e}_i$ as equal to the error e_i plus a deviation

$$\widehat{e}_i = Y_i - X_i'\widehat{\beta} = e_i - X_i'\left(\widehat{\beta} - \beta\right).$$

Thus the squared residual equals the squared error plus a deviation

$$\widehat{e}_i^2 = e_i^2 - 2e_i X_i'\left(\widehat{\beta} - \beta\right) + \left(\widehat{\beta} - \beta\right)' X_i X_i'\left(\widehat{\beta} - \beta\right). \tag{7.17}$$

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. This average is:

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n e_i^2 - 2\left(\frac{1}{n}\sum_{i=1}^n e_i X_i'\right)\left(\widehat{\beta} - \beta\right) + \left(\widehat{\beta} - \beta\right)'\left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)\left(\widehat{\beta} - \beta\right). \tag{7.18}$$

The WLLN implies that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n e_i^2 &\longrightarrow_p \sigma^2 \\ \frac{1}{n}\sum_{i=1}^n e_i X_i' &\longrightarrow_p \mathbb{E}\left[eX'\right] = 0 \\ \frac{1}{n}\sum_{i=1}^n X_i X_i' &\longrightarrow_p \mathbb{E}\left[XX'\right] = \boldsymbol{Q}_{XX}. \end{aligned}$$

Theorem 7.1 shows that $\widehat{\beta} \longrightarrow_p \beta$. Hence (7.18) converges in probability to σ^2 as desired.

Finally, since $n/(n-k) \to 1$ as $n \to \infty$ it follows that $s^2 = \left(\frac{n}{n-k}\right)\widehat{\sigma}^2 \longrightarrow_p \sigma^2$. Thus both estimators are consistent.

Theorem 7.4 Under Assumption 7.1, $\widehat{\sigma}^2 \xrightarrow[p]{} \sigma^2$ and $s^2 \xrightarrow[p]{} \sigma^2$ as $n \to \infty$.

7.6 Homoskedastic Covariance Matrix Estimation

Theorem 7.3 shows that $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ is asymptotically normal with asymptotic covariance matrix $\boldsymbol{V}_\beta$. For asymptotic inference (confidence intervals and tests) we need a consistent estimator of $\boldsymbol{V}_\beta$. Under homoskedasticity $\boldsymbol{V}_\beta$ simplifies to $\boldsymbol{V}_\beta^0 = \boldsymbol{Q}_{XX}^{-1}\sigma^2$ and in this section we consider the simplified problem of estimating $\boldsymbol{V}_\beta^0$.

The standard moment estimator of $\boldsymbol{Q}_{XX}$ is $\widehat{\boldsymbol{Q}}_{XX}$ defined in (7.1) and thus an estimator for $\boldsymbol{Q}_{XX}^{-1}$ is $\widehat{\boldsymbol{Q}}_{XX}^{-1}$. The standard estimator of σ^2 is the unbiased estimator s^2 defined in (4.31). Thus a natural plug-in estimator for $\boldsymbol{V}_\beta^0 = \boldsymbol{Q}_{XX}^{-1}\sigma^2$ is $\widehat{\boldsymbol{V}}_\beta^0 = \widehat{\boldsymbol{Q}}_{XX}^{-1}s^2$.

Consistency of $\widehat{\boldsymbol{V}}_\beta^0$ for $\boldsymbol{V}_\beta^0$ follows from consistency of the moment estimators $\widehat{\boldsymbol{Q}}_{XX}$ and s^2 and an application of the continuous mapping theorem. Specifically, Theorem 7.1 established $\widehat{\boldsymbol{Q}}_{XX} \xrightarrow[p]{} \boldsymbol{Q}_{XX}$, and Theorem 7.4 established $s^2 \xrightarrow[p]{} \sigma^2$. The function $\boldsymbol{V}_\beta^0 = \boldsymbol{Q}_{XX}^{-1}\sigma^2$ is a continuous function of $\boldsymbol{Q}_{XX}$ and σ^2 so long as $\boldsymbol{Q}_{XX} > 0$, which holds true under Assumption 7.1.4. It follows by the CMT that

$$\widehat{\boldsymbol{V}}_\beta^0 = \widehat{\boldsymbol{Q}}_{XX}^{-1}s^2 \xrightarrow[p]{} \boldsymbol{Q}_{XX}^{-1}\sigma^2 = \boldsymbol{V}_\beta^0$$

so that $\widehat{\boldsymbol{V}}_\beta^0$ is consistent for $\boldsymbol{V}_\beta^0$.

Theorem 7.5 Under Assumption 7.1, $\widehat{\boldsymbol{V}}_\beta^0 \xrightarrow[p]{} \boldsymbol{V}_\beta^0$ as $n \to \infty$.

It is instructive to notice that Theorem 7.5 does not require the assumption of homoskedasticity. That is, $\widehat{\boldsymbol{V}}_\beta^0$ is consistent for $\boldsymbol{V}_\beta^0$ regardless if the regression is homoskedastic or heteroskedastic. However, $\boldsymbol{V}_\beta^0 = \boldsymbol{V}_\beta = \operatorname{avar}\left[\widehat{\beta}\right]$ only under homoskedasticity. Thus, in the general case $\widehat{\boldsymbol{V}}_\beta^0$ is consistent for a well-defined but non-useful object.

7.7 Heteroskedastic Covariance Matrix Estimation

Theorems 7.3 established that the asymptotic covariance matrix of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ is $\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$. 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 $\boldsymbol{Q}_{XX}^{-1}$ is $\widehat{\boldsymbol{Q}}_{XX}^{-1}$ where $\widehat{\boldsymbol{Q}}_{XX}$ defined in (7.1).

The moment estimator for Ω is

$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2,$$

leading to the plug-in covariance matrix estimator

$$\widehat{\boldsymbol{V}}_\beta^{\mathrm{HC0}} = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\Omega}\widehat{\boldsymbol{Q}}_{XX}^{-1}. \tag{7.19}$$

You can check that $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$ where $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$ is the HC0 covariance matrix estimator from (4.36).

As shown in Theorem 7.1, $\widehat{\boldsymbol{Q}}_{XX}^{-1} \underset{p}{\longrightarrow} \boldsymbol{Q}_{XX}^{-1}$, so we just need to verify the consistency of $\widehat{\Omega}$. The key is to replace the squared residual $\widehat{e}_i^2$ with the squared error e_i^2, and then show that the difference is asymptotically negligible.

Specifically, observe that

$$\begin{aligned}\widehat{\Omega} &= \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2 + \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right).\end{aligned}$$

The first term is an average of the i.i.d. random variables $X_i X_i' e_i^2$, and therefore by the WLLN converges in probability to its expectation, namely,

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' e_i^2 \underset{p}{\longrightarrow} \mathbb{E}\left[XX' e^2\right] = \Omega.$$

Technically, this requires that Ω has finite elements, which was shown in (7.6).

To establish that $\widehat{\Omega}$ is consistent for Ω it remains to show that

$$\frac{1}{n}\sum_{i=1}^{n} X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right) \underset{p}{\longrightarrow} 0. \tag{7.20}$$

There are multiple ways to do this. A reasonably straightforward yet slightly tedious derivation is to start by applying the triangle inequality (B.16) using a matrix norm:

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| &\le \frac{1}{n}\sum_{i=1}^{n} \left\| X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| \\ &= \frac{1}{n}\sum_{i=1}^{n} \|X_i\|^2 \left|\widehat{e}_i^2 - e_i^2\right|. \end{aligned} \tag{7.21}$$

Then recalling the expression for the squared residual (7.17), apply the triangle inequality (B.1) and then the Schwarz inequality (B.12) twice

$$\begin{aligned}\left|\widehat{e}_i^2 - e_i^2\right| &\le 2\left|e_i X_i' \left(\widehat{\beta} - \beta\right)\right| + \left(\widehat{\beta} - \beta\right)' X_i X_i' \left(\widehat{\beta} - \beta\right) \\ &= 2|e_i| \left|X_i' \left(\widehat{\beta} - \beta\right)\right| + \left|\left(\widehat{\beta} - \beta\right)' X_i\right|^2 \\ &\le 2|e_i| \|X_i\| \left\|\widehat{\beta} - \beta\right\| + \|X_i\|^2 \left\|\widehat{\beta} - \beta\right\|^2. \end{aligned} \tag{7.22}$$

Combining (7.21) and (7.22), we find

$$\begin{aligned}\left\| \frac{1}{n}\sum_{i=1}^{n} X_i X_i' \left(\widehat{e}_i^2 - e_i^2\right) \right\| &\le 2\left(\frac{1}{n}\sum_{i=1}^{n} \|X_i\|^3 |e_i|\right) \left\|\widehat{\beta} - \beta\right\| + \left(\frac{1}{n}\sum_{i=1}^{n} \|X_i\|^4\right) \left\|\widehat{\beta} - \beta\right\|^2 \\ &= o_p(1). \end{aligned} \tag{7.23}$$

The expression is $o_p(1)$ because $\left\|\widehat{\beta} - \beta\right\| \underset{p}{\longrightarrow} 0$ and both averages in parenthesis are averages of random variables with finite expectation under Assumption 7.2 (and are thus $O_p(1)$). Indeed, by Hölder's inequality (B.31)

$$\mathbb{E}\left[\|X\|^3 |e|\right] \le \left(\mathbb{E}\left[\left(\|X\|^3\right)^{4/3}\right]\right)^{3/4} \left(\mathbb{E}\left[e^4\right]\right)^{1/4} = \left(\mathbb{E}\|X\|^4\right)^{3/4} \left(\mathbb{E}\left[e^4\right]\right)^{1/4} < \infty.$$

We have established (7.20) as desired.

Theorem 7.6 Under Assumption 7.2, as $n \to \infty$, $\widehat{\Omega} \xrightarrow[p]{} \Omega$ and $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} \xrightarrow[p]{} \boldsymbol{V}_{\beta}$.

For an alternative proof of this result, see Section 7.20.

7.8 Summary of Covariance Matrix Notation

The notation we have introduced may be somewhat confusing so it is helpful to write it down in one place.

The exact variance of $\widehat{\beta}$ (under the assumptions of the linear regression model) and the asymptotic variance of $\sqrt{n}\left(\widehat{\beta}-\beta\right)$ (under the more general assumptions of the linear projection model) are

$$\begin{aligned}\boldsymbol{V}_{\widehat{\beta}} &= \operatorname{var}\left[\widehat{\beta} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}'\boldsymbol{D}\boldsymbol{X}\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ \boldsymbol{V}_{\beta} &= \operatorname{avar}\left[\sqrt{n}\left(\widehat{\beta}-\beta\right)\right] = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}.\end{aligned}$$

The HC0 estimators of these two covariance matrices are

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ \widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} &= \widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\Omega}\widehat{\boldsymbol{Q}}_{XX}^{-1}\end{aligned}$$

and satisfy the simple relationship $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$.

Similarly, under the assumption of homoskedasticity the exact and asymptotic variances simplify to

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

Their standard estimators are

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} s^2 \\ \widehat{\boldsymbol{V}}_{\beta}^{0} &= \widehat{\boldsymbol{Q}}_{XX}^{-1} s^2\end{aligned}$$

which also satisfy the relationship $\widehat{\boldsymbol{V}}_{\beta}^{0} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{0}$.

The exact formula and estimators are useful when constructing test statistics and standard errors. However, for theoretical purposes the asymptotic formula (variances and their estimates) are more useful as these retain non-generate limits as the sample sizes diverge. That is why both sets of notation are useful.

7.9 Alternative Covariance Matrix Estimators*

In Section 7.7 we introduced $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}}$ as an estimator of $\boldsymbol{V}_{\beta}$. $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC0}}$ is a scaled version of $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC0}}$ from Section 4.14, where we also introduced the alternative HC1, HC2, and HC3 heteroskedasticity-robust covariance matrix estimators. We now discuss the consistency properties of these estimators.

To do so we introduce their scaled versions, e.g. $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC1}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC1}}$, $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC2}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC2}}$, and $\widehat{\boldsymbol{V}}_{\beta}^{\mathrm{HC3}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC3}}$. These are (alternative) estimators of the asymptotic covariance matrix $\boldsymbol{V}_{\beta}$.

First, consider $\widehat{\boldsymbol{V}}_\beta^{\text{HC1}}$. Notice that $\widehat{\boldsymbol{V}}_\beta^{\text{HC1}} = n\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\text{HC1}} = \frac{n}{n-k}\widehat{\boldsymbol{V}}_\beta^{\text{HC0}}$ where $\widehat{\boldsymbol{V}}_\beta^{\text{HC0}}$ was defined in (7.19) and shown consistent for $\boldsymbol{V}_\beta$ in Theorem 7.6. If k is fixed as $n \to \infty$, then $\frac{n}{n-k} \to 1$ and thus

$$\widehat{\boldsymbol{V}}_\beta^{\text{HC1}} = (1 + o(1))\widehat{\boldsymbol{V}}_\beta^{\text{HC0}} \xrightarrow[p]{} \boldsymbol{V}_\beta.$$

Thus $\widehat{\boldsymbol{V}}_\beta^{\text{HC1}}$ is consistent for $\boldsymbol{V}_\beta$.

The alternative estimators $\widehat{\boldsymbol{V}}_\beta^{\text{HC2}}$ and $\widehat{\boldsymbol{V}}_\beta^{\text{HC3}}$ take the form (7.19) but with $\widehat{\Omega}$ replaced by

$$\widetilde{\Omega} = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-2} X_i X_i' \widehat{e}_i^2$$

and

$$\overline{\Omega} = \frac{1}{n}\sum_{i=1}^{n}(1-h_{ii})^{-1} X_i X_i' \widehat{e}_i^2,$$

respectively. To show that these estimators also consistent for $\boldsymbol{V}_\beta$ given $\widehat{\Omega} \xrightarrow[p]{} \Omega$ it is sufficient to show that the differences $\widetilde{\Omega} - \widehat{\Omega}$ and $\overline{\Omega} - \widehat{\Omega}$ converge in probability to zero as $n \to \infty$.

The trick is the fact that the leverage values are asymptotically negligible:

$$h_n^* = \max_{1\le i\le n} h_{ii} = o_p(1). \tag{7.24}$$

(See Theorem 7.17 in Section 7.21.) Then using the triangle inequality (B.16)

$$\begin{aligned}\left\|\overline{\Omega} - \widehat{\Omega}\right\| &\le \frac{1}{n}\sum_{i=1}^{n}\left\|X_i X_i'\right\| \widehat{e}_i^2 \left|(1-h_{ii})^{-1} - 1\right| \\ &\le \left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2 \widehat{e}_i^2\right)\left|\left(1-h_n^*\right)^{-1} - 1\right|.\end{aligned}$$

The sum in parenthesis can be shown to be $O_p(1)$ under Assumption 7.2 by the same argument as in in the proof of Theorem 7.6. (In fact, it can be shown to converge in probability to $\mathbb{E}\left[\|X\|^2 e^2\right]$.) The term in absolute values is $o_p(1)$ by (7.24). Thus the product is $o_p(1)$ which means that $\overline{\Omega} = \widehat{\Omega} + o_p(1) \xrightarrow[p]{} \Omega$.

Similarly,

$$\begin{aligned}\left\|\widetilde{\Omega} - \widehat{\Omega}\right\| &\le \frac{1}{n}\sum_{i=1}^{n}\left\|X_i X_i'\right\| \widehat{e}_i^2 \left|(1-h_{ii})^{-2} - 1\right| \\ &\le \left(\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^2 \widehat{e}_i^2\right)\left|\left(1-h_n^*\right)^{-2} - 1\right| \\ &= o_p(1).\end{aligned}$$

Theorem 7.7 Under Assumption 7.2, as $n \to \infty$, $\widetilde{\Omega} \xrightarrow[p]{} \Omega$, $\overline{\Omega} \xrightarrow[p]{} \Omega$, $\widehat{\boldsymbol{V}}_\beta^{\text{HC1}} \xrightarrow[p]{} \boldsymbol{V}_\beta$, $\widehat{\boldsymbol{V}}_\beta^{\text{HC2}} \xrightarrow[p]{} \boldsymbol{V}_\beta$, and $\widehat{\boldsymbol{V}}_\beta^{\text{HC3}} \xrightarrow[p]{} \boldsymbol{V}_\beta$.

Theorem 7.7 shows that the alternative covariance matrix estimators are also consistent for the asymptotic covariance matrix.

To simplify notation, for the remainder of the chapter we will use the notation $\widehat{\boldsymbol{V}}_\beta$ and $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ to refer to any of the heteroskedasticity-consistent covariance matrix estimators HC0, HC1, HC2, and HC3, as they all have the same asymptotic limits.

7.10 Functions of Parameters

In most serious applications a researcher is actually interested in a specific transformation of the coefficient vector $\beta = (\beta_1, ..., \beta_k)$. For example, the researcher 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 θ as a function of the coefficients, e.g. $\theta = r(\beta)$ for some function $r : \mathbb{R}^k \to \mathbb{R}^q$. The estimate of θ is

$$\widehat{\theta} = r(\widehat{\beta}).$$

By the continuous mapping theorem (Theorem 6.6) and the fact $\widehat{\beta} \xrightarrow[p]{} \beta$ we can deduce that $\widehat{\theta}$ is consistent for θ if the function $r(\cdot)$ is continuous.

Theorem 7.8 Under Assumption 7.1, if $r(\beta)$ is continuous at the true value of β then as $n \to \infty$, $\widehat{\theta} \xrightarrow[p]{} \theta$.

Furthermore, if the transformation is sufficiently smooth, by the Delta Method (Theorem 6.8) we can show that $\widehat{\theta}$ is asymptotically normal.

Assumption 7.3 $r(\beta) : \mathbb{R}^k \to \mathbb{R}^q$ is continuously differentiable at the true value of β and $\boldsymbol{R} = \frac{\partial}{\partial \beta} r(\beta)'$ has rank q.

Theorem 7.9 Asymptotic Distribution of Functions of Parameters
Under Assumptions 7.2 and 7.3, as $n \to \infty$,

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow[d]{} \mathrm{N}(0, \boldsymbol{V}_\theta) \tag{7.25}$$

where $\boldsymbol{V}_\theta = \boldsymbol{R}' \boldsymbol{V}_\beta \boldsymbol{R}$.

In many cases the function $r(\beta)$ is linear:

$$r(\beta) = \boldsymbol{R}'\beta$$

for some $k \times q$ matrix $\boldsymbol{R}$. In particular if $\boldsymbol{R}$ is a "selector matrix"

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{I} \\ 0 \end{pmatrix}$$

then we can partition $\beta = (\beta_1', \beta_2')'$ so that $\boldsymbol{R}'\beta = \beta_1$. Then

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \boldsymbol{I} & 0 \end{pmatrix} \boldsymbol{V}_\beta \begin{pmatrix} \boldsymbol{I} \\ 0 \end{pmatrix} = \boldsymbol{V}_{11},$$

the upper-left sub-matrix of $\boldsymbol{V}_{11}$ given in (7.14). In this case (7.25) states that

$$\sqrt{n}\left(\widehat{\beta}_1 - \beta_1\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_{11}\right).$$

That is, subsets of $\widehat{\beta}$ are approximately normal with variances given by the conformable subcomponents of $\boldsymbol{V}$.

To illustrate the case of a nonlinear transformation take the example $\theta = \beta_j/\beta_l$ for $j \neq l$. Then

$$\boldsymbol{R} = \frac{\partial}{\partial \beta} r(\beta) = \begin{pmatrix} \frac{\partial}{\partial \beta_1}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_j}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_\ell}\left(\beta_j/\beta_l\right) \\ \vdots \\ \frac{\partial}{\partial \beta_k}\left(\beta_j/\beta_l\right) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 1/\beta_l \\ \vdots \\ -\beta_j/\beta_l^2 \\ \vdots \\ 0 \end{pmatrix} \tag{7.26}$$

so

$$\boldsymbol{V}_\theta = \boldsymbol{V}_{jj}/\beta_l^2 + \boldsymbol{V}_{ll}\beta_j^2/\beta_l^4 - 2\boldsymbol{V}_{jl}\beta_j/\beta_l^3$$

where $\boldsymbol{V}_{ab}$ denotes the ab^{th} element of $\boldsymbol{V}_\beta$.

For inference we need an estimator of the asymptotic covariance matrix $\boldsymbol{V}_\theta = \boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R}$. For this it is typical to use the plug-in estimator

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \beta} r(\widehat{\beta})'. \tag{7.27}$$

The derivative in (7.27) may be calculated analytically or numerically. By analytically, we mean working out the formula for the derivative and replacing the unknowns by point estimates. For example, if $\theta = \beta_j/\beta_l$ then $\frac{\partial}{\partial \beta} r(\beta)$ is (7.26). However in some cases the function $r(\beta)$ may be extremely complicated and a formula for the analytic derivative may not be easily available. In this case 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. The jl^{th} element of a numerical derivative $\widehat{\boldsymbol{R}}$ is

$$\widehat{\boldsymbol{R}}_{jl} = \frac{r_j(\widehat{\beta} + \delta_l \epsilon) - r_j(\widehat{\beta})}{\epsilon}$$

for some small ϵ.

The estimator of $\boldsymbol{V}_\theta$ is

$$\widehat{\boldsymbol{V}}_\theta = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}. \tag{7.28}$$

Alternatively, the homoskedastic covariance matrix estimator could be used leading to a homoskedastic covariance matrix estimator for θ.

$$\widehat{\boldsymbol{V}}_\theta^0 = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta^0\widehat{\boldsymbol{R}} = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\widehat{\boldsymbol{R}}s^2. \tag{7.29}$$

Given (7.27), (7.28) and (7.29) are simple to calculate using matrix operations.

As the primary justification for $\widehat{\boldsymbol{V}}_\theta$ is the asymptotic approximation (7.25), $\widehat{\boldsymbol{V}}_\theta$ is often called an **asymptotic covariance matrix estimator**.

The estimator $\widehat{\boldsymbol{V}}_\theta$ is consistent for $\boldsymbol{V}_\theta$ under the conditions of Theorem 7.9 because $\widehat{\boldsymbol{V}}_\beta \underset{p}{\longrightarrow} \boldsymbol{V}_\beta$ by Theorem 7.6 and

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \beta} r(\widehat{\beta})' \underset{p}{\longrightarrow} \frac{\partial}{\partial \beta} r(\beta)' = \boldsymbol{R}$$

because $\widehat{\beta} \underset{p}{\longrightarrow} \beta$ and the function $\frac{\partial}{\partial \beta} r(\beta)'$ is continuous in β.

Theorem 7.10 Under Assumptions 7.2 and 7.3, as $n \to \infty$, $\widehat{\boldsymbol{V}}_\theta \xrightarrow[p]{} \boldsymbol{V}_\theta$.

Theorem 7.10 shows that $\widehat{\boldsymbol{V}}_\theta$ is consistent for $\boldsymbol{V}_\theta$ and thus may be used for asymptotic inference. In practice we may set

$$\widehat{\boldsymbol{V}}_{\widehat{\theta}} = \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\beta}} \widehat{\boldsymbol{R}} = n^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}} \tag{7.30}$$

as an estimator of the variance of $\widehat{\theta}$.

7.11 Asymptotic Standard Errors

As described in Section 4.15, a standard error is an estimator of the standard deviation of the distribution of an estimator. Thus if $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ is an estimator of the covariance matrix of $\widehat{\beta}$ then standard errors are the square roots of the diagonal elements of this matrix. These take the form

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

Standard errors for $\widehat{\theta}$ are constructed similarly. Supposing that $\theta = h(\beta)$ is real-valued then the standard error for $\widehat{\theta}$ is the square root of (7.30)

$$s(\widehat{\theta}) = \sqrt{\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\beta}} \widehat{\boldsymbol{R}}} = \sqrt{n^{-1} \widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}}}.$$

When the justification is based on asymptotic theory we call $s(\widehat{\beta}_j)$ or $s(\widehat{\theta})$ an **asymptotic standard error** for $\widehat{\beta}_j$ or $\widehat{\theta}$. When reporting your results it is good practice to report standard errors for each reported estimate and this includes functions and transformations of your parameter estimates. This helps users of the work (including yourself) assess the estimation precision.

We illustrate using the log wage regression

$$\log(\mathit{wage}) = \beta_1 \, \mathit{education} + \beta_2 \, \mathit{experience} + \beta_3 \, \mathit{experience}^2/100 + \beta_4 + e.$$

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(wage) 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(wage) 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

$$\widehat{\log(wage)} = \underset{(0.008)}{0.118}\ education + \underset{(0.006)}{0.016}\ experience - \underset{(0.012)}{0.022}\ experience^2/100 + \underset{(0.157)}{0.947}\ . \tag{7.31}$$

The standard errors are the square roots of the HC2 covariance matrix estimate

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}} = \begin{pmatrix} 0.632 & 0.131 & -0.143 & -11.1 \\ 0.131 & 0.390 & -0.731 & -6.25 \\ -0.143 & -0.731 & 1.48 & 9.43 \\ -11.1 & -6.25 & 9.43 & 246 \end{pmatrix} \times 10^{-4}. \tag{7.32}$$

We calculate that

$$\widehat{\theta}_1 = 100\widehat{\beta}_1 = 100 \times 0.118 = 11.8$$

$$s(\widehat{\theta}_1) = \sqrt{100^2 \times 0.632 \times 10^{-4}} = 0.8$$

$$\widehat{\theta}_2 = 100\widehat{\beta}_2 + 20\widehat{\beta}_3 = 100 \times 0.016 - 20 \times 0.022 = 1.16$$

$$s(\widehat{\theta}_2) = \sqrt{\begin{pmatrix} 100 & 20 \end{pmatrix}\begin{pmatrix} 0.390 & -0.731 \\ -0.731 & 1.48 \end{pmatrix}\begin{pmatrix} 100 \\ 20 \end{pmatrix} \times 10^{-4}} = 0.55$$

$$\widehat{\theta}_3 = -50\widehat{\beta}_2/\widehat{\beta}_3 = 50 \times 0.016/0.022 = 35.2$$

$$s(\widehat{\theta}_3) = \sqrt{\begin{pmatrix} -50/\widehat{\beta}_3 & 50\widehat{\beta}_2/\widehat{\beta}_3^2 \end{pmatrix}\begin{pmatrix} 0.390 & -0.731 \\ -0.731 & 1.48 \end{pmatrix}\begin{pmatrix} -50/\widehat{\beta}_3 \\ 50\widehat{\beta}_2/\widehat{\beta}_3^2 \end{pmatrix} \times 10^{-4}} = 7.0.$$

The calculations show that the estimate of the percentage return to education is 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. The estimate of the experience level which maximizes expected log wages is 35 years with a standard error of 7.

In Stata the `nlcom` command can be used after estimation to perform the same calculations. To illustrate, after estimation of (7.31) use the commands given below. In each case, Stata reports the coefficient estimate, asymptotic standard error, and 95% confidence interval.

Stata Commands

```
nlcom 100*_b[education]
nlcom 100*_b[experience]+20*_b[exp2]
nlcom -50*_b[experience]/_b[exp2]
```

7.12 t-statistic

Let $\theta = r(\beta) : \mathbb{R}^k \to \mathbb{R}$ be a parameter of interest, $\widehat{\theta}$ its estimator, and $s(\widehat{\theta})$ its asymptotic standard error. Consider the statistic

$$T(\theta) = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})}. \tag{7.33}$$

Different writers call (7.33) a **t-statistic**, a **t-ratio**, a **z-statistic**, or a **studentized statistic**, sometimes using the different labels to distinguish between finite-sample and asymptotic inference. As the statistics themselves are always (7.33) we won't make this distinction, and will simply refer to $T(\theta)$ as a t-statistic or a t-ratio. We also often suppress the parameter dependence, writing it as T. The t-statistic is a function of the estimator, its standard error, and the parameter.

By Theorems 7.9 and 7.10, $\sqrt{n}\left(\widehat{\theta} - \theta\right) \underset{d}{\longrightarrow} \mathrm{N}(0, V_\theta)$ and $\widehat{V}_\theta \underset{p}{\longrightarrow} V_\theta$. Thus

$$\begin{aligned} T(\theta) &= \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \\ &= \frac{\sqrt{n}\left(\widehat{\theta} - \theta\right)}{\sqrt{\widehat{V}_\theta}} \\ &\underset{d}{\longrightarrow} \frac{\mathrm{N}(0, V_\theta)}{\sqrt{V_\theta}} \\ &= Z \sim \mathrm{N}(0,1). \end{aligned}$$

The last equality is the property that affine functions of normal variables are normal (Theorem 5.2).

This calculation requires that $V_\theta > 0$, otherwise the continuous mapping theorem cannot be employed. In practice this is an innocuous requirement as it only excludes degenerate sampling distributions. Formally we add the following assumption.

Assumption 7.4 $\boldsymbol{V}_\theta = \boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R} > 0$.

Assumption 7.4 states that $\boldsymbol{V}_\theta$ is positive definite. Since $\boldsymbol{R}$ is full rank under Assumption 7.3 a sufficient condition is that $\boldsymbol{V}_\beta > 0$. Since $\boldsymbol{Q}_{XX} > 0$ a sufficient condition is $\Omega > 0$. Thus Assumption 7.4 could be replaced by the assumption $\Omega > 0$. Assumption 7.4 is weaker so this is what we use.

Thus the asymptotic distribution of the t-ratio $T(\theta)$ is standard normal. Since this distribution does not depend on the parameters we say that $T(\theta)$ is **asymptotically pivotal**. In finite samples $T(\theta)$ is not necessarily pivotal but the property means that the dependence on unknowns diminishes as n increases.

It is also useful to consider the distribution of the **absolute t-ratio** $|T(\theta)|$. Since $T(\theta) \underset{d}{\longrightarrow} Z$ the continuous mapping theorem yields $|T(\theta)| \underset{d}{\longrightarrow} |Z|$. Letting $\Phi(u) = \mathbb{P}[Z \leq u]$ denote the standard normal distribution function we calculate that the distribution of $|Z|$ is

$$\begin{aligned}\mathbb{P}[|Z| \leq u] &= \mathbb{P}[-u \leq Z \leq u] \\ &= \mathbb{P}[Z \leq u] - \mathbb{P}[Z < -u] \\ &= \Phi(u) - \Phi(-u) \\ &= 2\Phi(u) - 1. \end{aligned} \tag{7.34}$$

Theorem 7.11 Under Assumptions 7.2, 7.3, and 7.4, $T(\theta) \underset{d}{\longrightarrow} Z \sim \mathrm{N}(0,1)$ and $|T(\theta)| \underset{d}{\longrightarrow} |Z|$.

The asymptotic normality of Theorem 7.11 is used to justify confidence intervals and tests for the parameters.

7.13 Confidence Intervals

The estimator $\widehat{\theta}$ is a **point estimator** for θ, meaning that $\widehat{\theta}$ is a single value in $\mathbb{R}^q$. A broader concept is a **set estimator** $\widehat{C}$ which is a collection of values in $\mathbb{R}^q$. When the parameter θ is real-valued then it is common to focus on sets of the form $\widehat{C} = [\widehat{L}, \widehat{U}]$ which is called an **interval estimator** for θ.

An interval estimator $\widehat{C}$ is a function of the data and hence is random. The **coverage probability** of the interval $\widehat{C} = [\widehat{L}, \widehat{U}]$ is $\mathbb{P}\left[\theta \in \widehat{C}\right]$. The randomness comes from $\widehat{C}$ as the parameter θ is treated as fixed. In Section 5.10 we introduced confidence intervals for the normal regression model which used the finite sample distribution of the t-statistic. When we are outside the normal regression model we cannot rely on the exact normal distribution theory but instead use asymptotic approximations. A benefit is that we can construct confidence intervals for general parameters of interest θ not just regression coefficients.

An interval estimator $\widehat{C}$ is called a **confidence interval** when the goal is to set the coverage probability to equal a pre-specified target such as 90% or 95%. $\widehat{C}$ is called a $1-\alpha$ confidence interval if $\inf_\theta \mathbb{P}_\theta\left[\theta \in \widehat{C}\right] = 1-\alpha$.

When $\widehat{\theta}$ is asymptotically normal with standard error $s(\widehat{\theta})$ the conventional confidence interval for θ takes the form

$$\widehat{C} = \left[\widehat{\theta} - c \times s(\widehat{\theta}), \quad \widehat{\theta} + c \times s(\widehat{\theta})\right] \tag{7.35}$$

where c equals the $1-\alpha$ quantile of the distribution of $|Z|$. Using (7.34) we calculate that c is equivalently the $1-\alpha/2$ quantile of the standard normal distribution. Thus, c solves

$$2\Phi(c) - 1 = 1 - \alpha.$$

This can be computed by, for example, `norminv(1-α/2)` in MATLAB. The confidence interval (7.35) is symmetric about the point estimator $\widehat{\theta}$ and its length is proportional to the standard error $s(\widehat{\theta})$.

Equivalently, (7.35) is the set of parameter values for θ such that the t-statistic $T(\theta)$ is smaller (in absolute value) than c, that is

$$\widehat{C} = \{\theta : |T(\theta)| \leq c\} = \left\{\theta : -c \leq \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \leq c\right\}.$$

The coverage probability of this confidence interval is

$$\mathbb{P}\left[\theta \in \widehat{C}\right] = \mathbb{P}\left[|T(\theta)| \leq c\right] \to \mathbb{P}\left[|Z| \leq c\right] = 1 - \alpha$$

where the limit is taken as $n \to \infty$, and holds because $T(\theta)$ is asymptotically $|Z|$ by Theorem 7.11. We call the limit the **asymptotic coverage probability** and call $\widehat{C}$ an asymptotic $1-\alpha$% confidence interval for θ. Since the t-ratio is asymptotically pivotal the asymptotic coverage probability is independent of the parameter θ.

It is useful to contrast the confidence interval (7.35) with (5.8) for the normal regression model. They are similar but there are differences. The normal regression interval (5.8) only applies to regression coefficients β not to functions θ of the coefficients. The normal interval (5.8) also is constructed with the homoskedastic standard error, while (7.35) can be constructed with a heteroskedastic-robust standard error. Furthermore, the constants c in (5.8) are calculated using the student t distribution, while c in (7.35) are calculated using the normal distribution. The difference between the student t and normal values are typically small in practice (since sample sizes are large in typical economic applications). However, since the student t values are larger it results in slightly larger confidence intervals which is reasonable. (A practical rule of thumb is that if the sample sizes are sufficiently small that it makes a difference then neither (5.8) nor (7.35) should be trusted.) Despite these differences the coincidence of the intervals means that inference on regression coefficients is generally robust to using either the exact normal sampling assumption or the asymptotic large sample approximation, at least in large samples.

Stata by default reports 95% confidence intervals for each coefficient where the critical values c are calculated using the t_{n-k} distribution. This is done for all standard error methods even though it is only exact for homoskedastic standard errors and under normality.

The standard coverage probability for confidence intervals is 95%, leading to the choice $c = 1.96$ for the constant in (7.35). Rounding 1.96 to 2, we obtain the most commonly used confidence interval in applied econometric practice

$$\widehat{C} = \left[\widehat{\theta} - 2s(\widehat{\theta}),\quad \widehat{\theta} + 2s(\widehat{\theta})\right].$$

This is a useful rule-of thumb. This asymptotic 95% confidence interval $\widehat{C}$ is simple to compute and can be roughly calculated from tables of coefficient estimates and standard errors. (Technically, it is an asymptotic 95.4% interval due to the substitution of 2.0 for 1.96 but this distinction is overly precise.)

Theorem 7.12 Under Assumptions 7.2, 7.3 and 7.4, for $\widehat{C}$ defined in (7.35) with $c = \Phi^{-1}(1-\alpha/2)$, $\mathbb{P}\left[\theta \in \widehat{C}\right] \to 1-\alpha$. For $c = 1.96$, $\mathbb{P}\left[\theta \in \widehat{C}\right] \to 0.95$.

Confidence intervals are a simple yet effective tool to assess estimation uncertainty. When reading a set of empirical results look at the estimated coefficient estimates and the standard errors. For a parameter of interest compute the confidence interval $\widehat{C}$ and consider the meaning of the spread of the suggested values. If the range of values in the confidence interval are too wide to learn about θ then do not jump to a conclusion about θ based on the point estimate alone.

For illustration, consider the three examples presented in Section 7.11 based on the log wage regression for married Black women.

Percentage return to education. A 95% asymptotic confidence interval is $11.8 \pm 1.96 \times 0.8 = [10.2, 13.3]$. This is reasonably tight.

Percentage return to experience (per year) for individuals with 10 years experience. A 90% asymptotic confidence interval is $1.1 \pm 1.645 \times 0.4 = [0.5, 1.8]$. The interval is positive but broad. This indicates that the return to experience is positive, but of uncertain magnitude.

Experience level which maximizes expected log wages. An 80% asymptotic confidence interval is $35 \pm 1.28 \times 7 = [26, 44]$. This is rather imprecise, indicating that the estimates are not very informative regarding this parameter.

7.14 Regression Intervals

In the linear regression model the conditional expectation of Y given $X = x$ is

$$m(x) = \mathbb{E}[Y \mid X = x] = x'\beta.$$

In some cases we want to estimate $m(x)$ at a particular point x. Notice that this is a linear function of β. Letting $r(\beta) = x'\beta$ and $\theta = r(\beta)$ we see that $\widehat{m}(x) = \widehat{\theta} = x'\widehat{\beta}$ and $\boldsymbol{R} = x$ so $s(\widehat{\theta}) = \sqrt{x'\widehat{\boldsymbol{V}}_{\widehat{\beta}}x}$. Thus an asymptotic 95% confidence interval for $m(x)$ is

$$\left[x'\widehat{\beta} \pm 1.96\sqrt{x'\widehat{\boldsymbol{V}}_{\widehat{\beta}}x}\right].$$

It is interesting to observe that if this is viewed as a function of x the width of the confidence interval is dependent on x.

To illustrate we return to the log wage regression (3.12) of Section 3.7. The estimated regression equation is

$$\widehat{\log(wage)} = x'\widehat{\beta} = 0.155x + 0.698$$

where $x =$*education*. The covariance matrix estimate from (4.43) is

$$\widehat{\boldsymbol{V}}_{\widehat{\beta}} = \begin{pmatrix} 0.001 & -0.015 \\ -0.015 & 0.243 \end{pmatrix}.$$

Thus the 95% confidence interval for the regression is

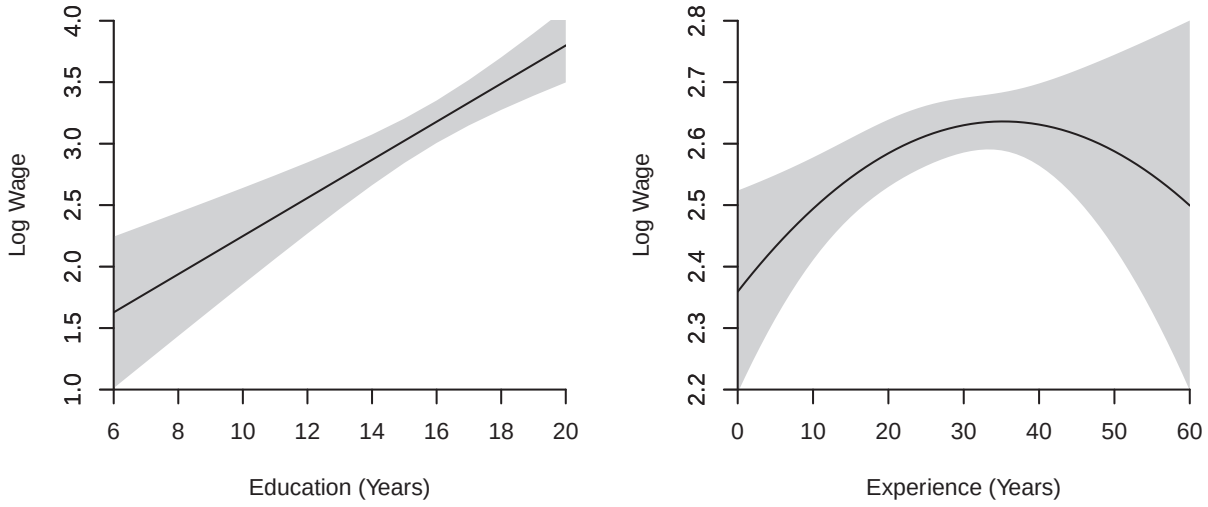
$$0.155x + 0.698 \pm 1.96\sqrt{0.001x^2 - 0.030x + 0.243}.$$

The estimated regression and 95% intervals are shown in Figure 7.4(a). Notice that the confidence bands take a hyperbolic shape. This means that the regression line is less precisely estimated for large and small values of *education*.

Plots of the estimated regression line and confidence intervals are especially useful when the regression includes nonlinear terms. To illustrate consider the log wage regression (7.31) which includes experience and its square and covariance matrix estimate (7.32). 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 fix *education* at a specific value. We select *education*=12. This only affects the level of the estimated regression since *education* enters without an interaction. Define the points of evaluation

$$z(x) = \begin{pmatrix} 12 \\ x \\ x^2/100 \\ 1 \end{pmatrix}$$

where $x =$*experience*.



(a) Wage on Education

(b) Wage on Experience

Figure 7.4: Regression Intervals

The 95% regression interval for *education*=12 as a function of x =*experience* is

$$
\begin{aligned}
&0.118 \times 12 \ + 0.016\, x - 0.022\, x^2/100 + 0.947 \\
&\pm 1.96 \sqrt{z(x)' \begin{pmatrix} 0.632 & 0.131 & -0.143 & -11.1 \\ 0.131 & 0.390 & -0.731 & -6.25 \\ -0.143 & -0.731 & 1.48 & 9.43 \\ -11.1 & -6.25 & 9.43 & 246 \end{pmatrix} z(x) \times 10^{-4}} \\
&= 0.016\, x - .00022\, x^2 + 2.36 \\
&\pm 0.0196 \sqrt{70.608 - 9.356\, x + 0.54428\, x^2 - 0.01462\, x^3 + 0.000148\, x^4}.
\end{aligned}
$$

The estimated regression and 95% intervals are shown in Figure 7.4(b). The regression interval widens greatly for small and large values of experience indicating considerable uncertainty about the effect of experience on mean wages for this population. The confidence bands take a more complicated shape than in Figure 7.4(a) due to the nonlinear specification.

7.15 Forecast Intervals

Suppose we are given a value of the regressor vector 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 $X_{n+1} = x$ which will generally be a function of x. A reasonable forecasting rule is the conditional expectation $m(x)$ as it is the mean-square minimizing forecast. A point forecast is the estimated conditional expectation $\widehat{m}(x) = x'\widehat{\beta}$. We would also like a measure of uncertainty for the forecast.

The forecast error is $\widehat{e}_{n+1} = Y_{n+1} - \widehat{m}(x) = e_{n+1} - x'\left(\widehat{\beta} - \beta\right)$. As the out-of-sample error e_{n+1} is inde-

pendent of the in-sample estimator $\widehat{\beta}$ this has conditional variance

$$\begin{aligned}\mathbb{E}\left[\widehat{e}_{n+1}^2 \mid X_{n+1}=x\right] &= \mathbb{E}\left[e_{n+1}^2 - 2x'\left(\widehat{\beta}-\beta\right)e_{n+1} + x'\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'x \mid X_{n+1}=x\right] \\ &= \mathbb{E}\left[e_{n+1}^2 \mid X_{n+1}=x\right] + x'\mathbb{E}\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'\right]x \\ &= \sigma^2(x) + x'\boldsymbol{V}_{\widehat{\beta}}x. \end{aligned} \tag{7.36}$$

Under homoskedasticity, $\mathbb{E}\left[e_{n+1}^2 \mid X_{n+1}\right] = \sigma^2$. In this case a simple estimator of (7.36) is $\widehat{\sigma}^2 + x'\boldsymbol{V}_{\widehat{\beta}}x$ so a standard error for the forecast is $\widehat{s}(x) = \sqrt{\widehat{\sigma}^2 + x'\boldsymbol{V}_{\widehat{\beta}}x}$. Notice that this is different from the standard error for the conditional expectation.

The conventional 95% forecast interval for Y_{n+1} uses a normal approximation and equals $\left[x'\widehat{\beta} \pm 2\widehat{s}(x)\right]$. It is difficult, however, to fully justify this choice. It would be correct if we have a normal approximation to the ratio

$$\frac{e_{n+1} - x'\left(\widehat{\beta}-\beta\right)}{\widehat{s}(x)}.$$

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.

An accurate forecast interval would use the conditional distribution of e_{n+1} given $X_{n+1} = x$, which is more challenging to estimate. Due to this difficulty many applied forecasters use the simple approximate interval $\left[x'\widehat{\beta} \pm 2\widehat{s}(x)\right]$ despite the lack of a convincing justification.

7.16 Wald Statistic

Let $\theta = r(\beta) : \mathbb{R}^k \to \mathbb{R}^q$ be any parameter vector of interest, $\widehat{\theta}$ its estimator, and $\widehat{\boldsymbol{V}}_{\widehat{\theta}}$ its covariance matrix estimator. Consider the quadratic form

$$W(\theta) = \left(\widehat{\theta}-\theta\right)'\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{-1}\left(\widehat{\theta}-\theta\right) = n\left(\widehat{\theta}-\theta\right)'\widehat{\boldsymbol{V}}_{\theta}^{-1}\left(\widehat{\theta}-\theta\right). \tag{7.37}$$

where $\widehat{\boldsymbol{V}}_{\theta} = n\widehat{\boldsymbol{V}}_{\widehat{\theta}}$. When $q = 1$, then $W(\theta) = T(\theta)^2$ is the square of the t-ratio. When $q > 1$, $W(\theta)$ is typically called a **Wald statistic** as it was proposed by Wald (1943). We are interested in its sampling distribution.

The asymptotic distribution of $W(\theta)$ is simple to derive given Theorem 7.9 and Theorem 7.10. They show that $\sqrt{n}\left(\widehat{\theta}-\theta\right) \underset{d}{\longrightarrow} Z \sim \mathrm{N}(0, \boldsymbol{V}_{\theta})$ and $\widehat{\boldsymbol{V}}_{\theta} \underset{p}{\longrightarrow} \boldsymbol{V}_{\theta}$. It follows that

$$W(\theta) = \sqrt{n}\left(\widehat{\theta}-\theta\right)'\widehat{\boldsymbol{V}}_{\theta}^{-1}\sqrt{n}\left(\widehat{\theta}-\theta\right) \underset{d}{\longrightarrow} Z'\boldsymbol{V}_{\theta}^{-1}Z$$

a quadratic in the normal random vector Z. As shown in Theorem 5.3.5 the distribution of this quadratic form is χ_q^2, a chi-square random variable with q degrees of freedom.

Theorem 7.13 Under Assumptions 7.2, 7.3 and 7.4, as $n \to \infty$, $W(\theta) \underset{d}{\longrightarrow} \chi_q^2$.

Theorem 7.13 is used to justify multivariate confidence regions and multivariate hypothesis tests.

7.17 Homoskedastic Wald Statistic

Under the conditional homoskedasticity assumption $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$ we can construct the Wald statistic using the homoskedastic covariance matrix estimator $\widehat{\boldsymbol{V}}_\theta^0$ defined in (7.29). This yields a homoskedastic Wald statistic

$$W^0(\theta) = \left(\widehat{\theta} - \theta\right)' \left(\widehat{\boldsymbol{V}}_{\widehat{\theta}}^0\right)^{-1} \left(\widehat{\theta} - \theta\right) = n\left(\widehat{\theta} - \theta\right)' \left(\widehat{\boldsymbol{V}}_\theta^0\right)^{-1} \left(\widehat{\theta} - \theta\right). \tag{7.38}$$

Under the assumption of conditional homoskedasticity it has the same asymptotic distribution as $W(\theta)$.

Theorem 7.14 Under Assumptions 7.2, 7.3, and $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2 > 0$, as $n \to \infty$, $W^0(\theta) \xrightarrow[d]{} \chi_q^2$.

7.18 Confidence Regions

A confidence region $\widehat{C}$ is a set estimator for $\theta \in \mathbb{R}^q$ when $q > 1$. A confidence region $\widehat{C}$ is a set in $\mathbb{R}^q$ intended to cover the true parameter value with a pre-selected probability $1-\alpha$. Thus an ideal confidence region has the coverage probability $\mathbb{P}\left[\theta \in \widehat{C}\right] = 1 - \alpha$. In practice it is typically not possible to construct a region with exact coverage but we can calculate its asymptotic coverage.

When the parameter estimator satisfies the conditions of Theorem 7.13 a good choice for a confidence region is the ellipse

$$\widehat{C} = \{\theta : W(\theta) \leq c_{1-\alpha}\}$$

with $c_{1-\alpha}$ the $1-\alpha$ quantile of the χ_q^2 distribution. (Thus $F_q(c_{1-\alpha}) = 1-\alpha$.) It can be computed by, for example, `chi2inv(1-`α`,q)` in MATLAB.

Theorem 7.13 implies

$$\mathbb{P}\left[\theta \in \widehat{C}\right] \to \mathbb{P}\left[\chi_q^2 \leq c_{1-\alpha}\right] = 1 - \alpha$$

which shows that $\widehat{C}$ has asymptotic coverage $1-\alpha$.

To illustrate the construction of a confidence region, consider the estimated regression (7.31) of

$$\widehat{\log(\mathit{wage})} = \beta_1\, \mathit{education} + \beta_2\, \mathit{experience} + \beta_3\, \mathit{experience}^2/100 + \beta_4.$$

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

$$\widehat{\theta} = \begin{pmatrix} 100 & 0 & 0 & 0 \\ 0 & 100 & 20 & 0 \end{pmatrix} \widehat{\beta} = \begin{pmatrix} 11.8 \\ 1.2 \end{pmatrix},$$

and have the covariance matrix estimate

$$\widehat{\boldsymbol{V}}_{\widehat{\theta}} = \begin{pmatrix} 0 & 100 & 0 & 0 \\ 0 & 0 & 100 & 20 \end{pmatrix} \widehat{\boldsymbol{V}}_{\widehat{\beta}} \begin{pmatrix} 0 & 0 \\ 100 & 0 \\ 0 & 100 \\ 0 & 20 \end{pmatrix} = \begin{pmatrix} 0.632 & 0.103 \\ 0.103 & 0.157 \end{pmatrix}$$

with inverse

$$\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{-1} = \begin{pmatrix} 1.77 & -1.16 \\ -1.16 & 7.13 \end{pmatrix}.$$

Thus the Wald statistic is

$$\begin{aligned} W(\theta) &= \left(\widehat{\theta} - \theta\right)' \widehat{\boldsymbol{V}}_{\widehat{\theta}}^{-1} \left(\widehat{\theta} - \theta\right) \\ &= \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix}' \begin{pmatrix} 1.77 & -1.16 \\ -1.16 & 7.13 \end{pmatrix} \begin{pmatrix} 11.8 - \theta_1 \\ 1.2 - \theta_2 \end{pmatrix} \\ &= 1.77 \left(11.8 - \theta_1\right)^2 - 2.32 \left(11.8 - \theta_1\right)\left(1.2 - \theta_2\right) + 7.13 \left(1.2 - \theta_2\right)^2. \end{aligned}$$

The 90% quantile of the χ_2^2 distribution is 4.605 (we use the χ_2^2 distribution as the dimension of θ is two) so an asymptotic 90% confidence region for the two parameters is the interior of the ellipse $W(\theta) = 4.605$ which is displayed in Figure 7.5. Since the estimated correlation of the two coefficient estimates is modest (about 0.3) the region is modestly elliptical.

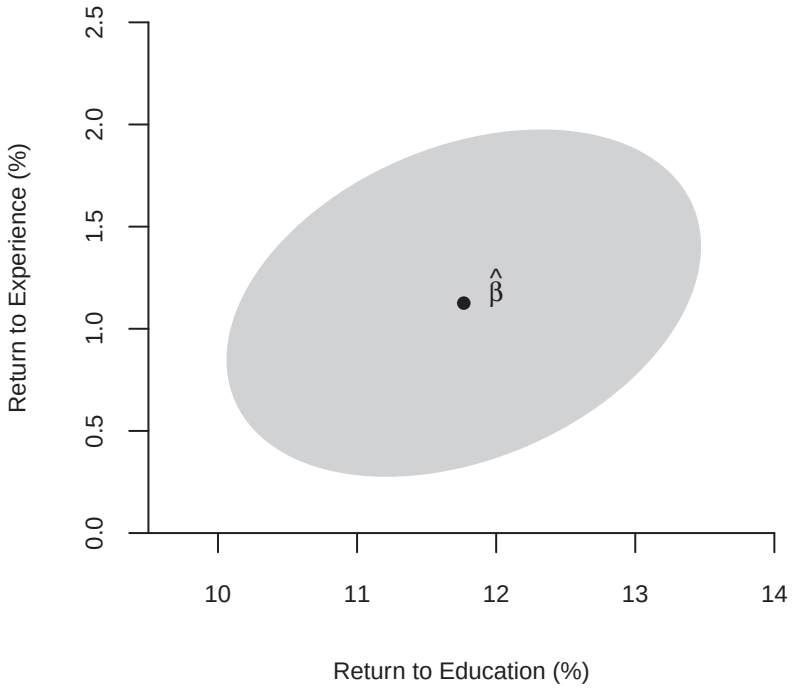


Figure 7.5: Confidence Region for Return to Experience and Return to Education

7.19 Edgeworth Expansion*

Theorem 7.11 showed that the t-ratio $T(\theta)$ is asymptotically normal. In practice this means that we use the normal distribution to approximate the finite sample distribution of T. How good is this approximation? Some insight into the accuracy of the normal approximation can be obtained by an

Edgeworth expansion which is a higher-order approximation to the distribution of T. The following result is an application of Theorem 9.11 of *Probability and Statistics for Economists*.

Theorem 7.15 Under Assumptions 7.2, 7.3, $\Omega > 0$, $\mathbb{E}\|e\|^{16} < \infty$, $\mathbb{E}\|X\|^{16} < \infty$, $g(\beta)$ has five continuous derivatives in a neighborhood of β, and $\mathbb{E}\left[\exp\left(t\left(\|e\|^4 + \|X\|^4\right)\right)\right] \leq B < 1$, as $n \to \infty$

$$\mathbb{P}[T(\theta) \leq x] = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + n^{-1} p_2(x)\phi(x) + o\left(n^{-1}\right)$$

uniformly in x, where $p_1(x)$ is an even polynomial of order 2 and $p_2(x)$ is an odd polynomial of degree 5 with coefficients depending on the moments of e and X up to order 16.

Theorem 7.15 shows that the finite sample distribution of the t-ratio can be approximated up to $o(n^{-1})$ by the sum of three terms, the first being the standard normal distribution, the second a $O\left(n^{-1/2}\right)$ adjustment, and the third a $O\left(n^{-1}\right)$ adjustment.

Consider a one-sided confidence interval $\widehat{C} = \left[\widehat{\theta} - z_{1-\alpha} s(\widehat{\theta}), \infty\right)$ where $z_{1-\alpha}$ is the $1-\alpha^{th}$ quantile of $Z \sim \mathrm{N}(0,1)$, thus $\Phi(z_{1-\alpha}) - 1 - \alpha$. Then

$$\begin{aligned}\mathbb{P}\left[\theta \in \widehat{C}\right] &= \mathbb{P}[T(\theta) \leq z_{1-\alpha}] \\ &= \Phi(z_{1-\alpha}) + n^{-1/2} p_1(z_{1-\alpha})\phi(z_{1-\alpha}) + O\left(n^{-1}\right) \\ &= 1 - \alpha + O\left(n^{-1/2}\right).\end{aligned}$$

This means that the actual coverage is within $O\left(n^{-1/2}\right)$ of the desired $1-\alpha$ level.

Now consider a two-sided interval $\widehat{C} = \left[\widehat{\theta} - z_{1-\alpha/2} s(\widehat{\theta}), \widehat{\theta} + z_{1-\alpha/2} s(\widehat{\theta})\right]$. It has coverage

$$\begin{aligned}\mathbb{P}\left[\theta \in \widehat{C}\right] &= \mathbb{P}[|T(\theta)| \leq z_{1-\alpha/2}] \\ &= 2\Phi(z_{1-\alpha/2}) - 1 + n^{-1} 2 p_2(z_{1-\alpha/2})\phi(z_{1-\alpha/2}) + o\left(n^{-1}\right) \\ &= 1 - \alpha + O\left(n^{-1}\right).\end{aligned}$$

This means that the actual coverage is within $O\left(n^{-1}\right)$ of the desired $1-\alpha$ level. The accuracy is better than the one-sided interval because the $O\left(n^{-1/2}\right)$ term in the Edgeworth expansion has offsetting effects in the two tails of the distribution.

7.20 Uniformly Consistent Residuals*

It seems natural to view the residuals $\widehat{e}_i$ as estimators of the unknown errors e_i. Are they consistent? In this section we develop a convergence result.

We can write the residual as

$$\widehat{e}_i = Y_i - X_i'\widehat{\beta} = e_i - X_i'\left(\widehat{\beta} - \beta\right). \tag{7.39}$$

Since $\widehat{\beta} - \beta \xrightarrow[p]{} 0$ it seems reasonable to guess that $\widehat{e}_i$ will be close to e_i if n is large.

We can bound the difference in (7.39) using the Schwarz inequality (B.12) to find

$$|\widehat{e}_i - e_i| = \left|X_i'\left(\widehat{\beta} - \beta\right)\right| \leq \|X_i\| \left\|\widehat{\beta} - \beta\right\|. \tag{7.40}$$

To bound (7.40) we can use $\left\|\widehat{\beta}-\beta\right\| = O_p\left(n^{-1/2}\right)$ from Theorem 7.3. We also need to bound the random variable $\|X_i\|$. If the regressor is bounded, that is, $\|X_i\| \le B < \infty$, then $|\widehat{e}_i - e_i| \le B\left\|\widehat{\beta}-\beta\right\| = O_p\left(n^{-1/2}\right)$. However if the regressor does not have bounded support then we have to be more careful.

The key is Theorem 6.15 which shows that $\mathbb{E}\|X\|^r < \infty$ implies $X_i = o_p\left(n^{1/r}\right)$ uniformly in i, or

$$n^{-1/r}\max_{1\le i\le n}\|X_i\| \xrightarrow[p]{} 0.$$

Applied to (7.40) we obtain

$$\max_{1\le i\le n}|\widehat{e}_i - e_i| \le \max_{1\le i\le n}\|X_i\|\left\|\widehat{\beta}-\beta\right\| = o_p\left(n^{-1/2+1/r}\right).$$

We have shown the following.

Theorem 7.16 Under Assumption 7.2 and $\mathbb{E}\|X\|^r < \infty$, then

$$\max_{1\le i\le n}|\widehat{e}_i - e_i| = o_p\left(n^{-1/2+1/r}\right). \tag{7.41}$$

The rate of convergence in (7.41) depends on r. Assumption 7.2 requires $r \ge 4$ so the rate of convergence is at least $o_p\left(n^{-1/4}\right)$. As r increases the rate improves.

We mentioned in Section 7.7 that there are multiple ways to prove the consistency of the covariance matrix estimator $\widehat{\Omega}$. We now show that Theorem 7.16 provides one simple method to establish (7.23) and thus Theorem 7.6. Let $q_n = \max_{1\le i\le n}|\widehat{e}_i - e_i| = o_p\left(n^{-1/4}\right)$. Since $\widehat{e}_i^2 - e_i^2 = 2e_i\left(\widehat{e}_i - e_i\right) + \left(\widehat{e}_i - e_i\right)^2$, then

$$\begin{aligned}
\left\|\frac{1}{n}\sum_{i=1}^n X_i X_i'\left(\widehat{e}_i^2 - e_i^2\right)\right\| &\le \frac{1}{n}\sum_{i=1}^n \left\|X_i X_i'\right\|\left|\widehat{e}_i^2 - e_i^2\right| \\
&\le \frac{2}{n}\sum_{i=1}^n \|X_i\|^2 |e_i|\,|\widehat{e}_i - e_i| + \frac{1}{n}\sum_{i=1}^n \|X_i\|^2 |\widehat{e}_i - e_i|^2 \\
&\le \frac{2}{n}\sum_{i=1}^n \|X_i\|^2 |e_i|\, q_n + \frac{1}{n}\sum_{i=1}^n \|X_i\|^2 q_n^2 \\
&\le o_p\left(n^{-1/4}\right).
\end{aligned}$$

7.21 Asymptotic Leverage*

Recall the definition of leverage from (3.40) $h_{ii} = X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}X_i$. These are the diagonal elements of the projection matrix $\boldsymbol{P}$ and appear in the formula for leave-one-out prediction errors and HC2 and HC3 covariance matrix estimators. We can show that under i.i.d. 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 $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \underset{p}{\longrightarrow} \boldsymbol{Q}_{XX} > 0$, by the CMT $\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right) \underset{p}{\longrightarrow} \lambda_{\min}\left(\boldsymbol{Q}_{XX}\right) > 0$. (The latter is positive since $\boldsymbol{Q}_{XX}$ is positive definite and thus all its eigenvalues are positive.) Then by the Quadratic Inequality (B.18)

$$\begin{aligned} h_{ii} &= X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \\ &\leq \lambda_{\max}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\left(X_i'X_i\right) \\ &= \left(\lambda_{\min}\left(\frac{1}{n}\boldsymbol{X}'\boldsymbol{X}\right)\right)^{-1}\frac{1}{n}\left\|X_i\right\|^2 \\ &\leq \left(\lambda_{\min}\left(\boldsymbol{Q}_{XX}\right)+o_p(1)\right)^{-1}\frac{1}{n}\max_{1\leq i\leq n}\left\|X_i\right\|^2. \end{aligned} \tag{7.42}$$

Theorem 6.15 shows that $\mathbb{E}\left\|X\right\|^r < \infty$ implies $\max_{1\leq i\leq n}\left\|X_i\right\|^2 = \left(\max_{1\leq i\leq n}\left\|X_i\right\|\right)^2 = o_p\left(n^{2/r}\right)$ and thus (7.42) is $o_p\left(n^{2/r-1}\right)$.

Theorem 7.17 If X_i is i.i.d., $\boldsymbol{Q}_{XX} > 0$, and $\mathbb{E}\left\|X\right\|^r < \infty$ for some $r \geq 2$, then $\max_{1\leq i\leq n} h_{ii} = o_p\left(n^{2/r-1}\right)$.

For any $r \geq 2$ then $h_{ii} = o_p(1)$ (uniformly in $i \leq n$). Larger r implies a faster rate of convergence. For example $r = 4$ implies $h_{ii} = o_p\left(n^{-1/2}\right)$.

Theorem (7.17) 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.

7.22 Exercises

Exercise 7.1 Take the model $Y = X_1'\beta_1 + X_2'\beta_2 + e$ with $\mathbb{E}\left[Xe\right] = 0$. Suppose that β_1 is estimated by regressing Y on X_1 only. Find the probability limit of this estimator. In general, is it consistent for β_1? If not, under what conditions is this estimator consistent for β_1?

Exercise 7.2 Take the model $Y = X'\beta + e$ with $\mathbb{E}\left[Xe\right] = 0$. Define the **ridge regression** estimator

$$\widehat{\beta} = \left(\sum_{i=1}^{n} X_i X_i' + \lambda \boldsymbol{I}_k\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right) \tag{7.43}$$

here $\lambda > 0$ is a fixed constant. Find the probability limit of $\widehat{\beta}$ as $n \to \infty$. Is $\widehat{\beta}$ consistent for β?

Exercise 7.3 For the ridge regression estimator (7.43), set $\lambda = cn$ where $c > 0$ is fixed as $n \to \infty$. Find the probability limit of $\widehat{\beta}$ as $n \to \infty$.

Exercise 7.4 Verify some of the calculations reported in Section 7.4. Specifically, suppose that X_1 and X_2 only take the values $\{-1,+1\}$, symmetrically, with

$$\begin{aligned}\mathbb{P}[X_1 = X_2 = 1] = \mathbb{P}[X_1 = X_2 = -1] &= 3/8 \\ \mathbb{P}[X_1 = 1, X_2 = -1] = \mathbb{P}[X_1 = -1, X_{2i} = 1] &= 1/8 \\ \mathbb{E}\left[e_i^2 \mid X_1 = X_2\right] &= \frac{5}{4} \\ \mathbb{E}\left[e_i^2 \mid X_1 \neq X_2\right] &= \frac{1}{4}.\end{aligned}$$

Verify the following:

(a) $\mathbb{E}[X_1] = 0$

(b) $\mathbb{E}\left[X_1^2\right] = 1$

(c) $\mathbb{E}[X_1 X_2] = \frac{1}{2}$

(d) $\mathbb{E}\left[e^2\right] = 1$

(e) $\mathbb{E}\left[X_1^2 e^2\right] = 1$

(f) $\mathbb{E}\left[X_1 X_2 e^2\right] = \frac{7}{8}$.

Exercise 7.5 Show (7.13)-(7.16).

Exercise 7.6 The model is

$$\begin{aligned}Y &= X'\beta + e \\ \mathbb{E}[Xe] &= 0 \\ \Omega &= \mathbb{E}\left[XX'e^2\right].\end{aligned}$$

Find the method of moments estimators $(\widehat{\beta}, \widehat{\Omega})$ for (β, Ω).

Exercise 7.7 Of the variables (Y^*, Y, X) only the pair (Y, X) are observed. In this case we say that Y^* is a **latent variable**. Suppose

$$\begin{aligned}Y^* &= X'\beta + e \\ \mathbb{E}[Xe] &= 0 \\ Y &= Y^* + u\end{aligned}$$

where u is a measurement error satisfying

$$\begin{aligned}\mathbb{E}[Xu] &= 0 \\ \mathbb{E}\left[Y^* u\right] &= 0.\end{aligned}$$

Let $\widehat{\beta}$ denote the OLS coefficient from the regression of Y on X.

(a) Is β the coefficient from the linear projection of Y on X?

(b) Is $\widehat{\beta}$ consistent for β as $n \to \infty$?

(c) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ as $n \to \infty$.

Exercise 7.8 Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\sigma}^2 - \sigma^2\right)$ as $n \to \infty$.

Exercise 7.9 The model is $Y = X\beta + e$ with $\mathbb{E}[e \mid X] = 0$ and $X \in \mathbb{R}$. Consider the two estimators

$$\widehat{\beta} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$
$$\widetilde{\beta} = \frac{1}{n}\sum_{i=1}^n \frac{Y_i}{X_i}.$$

(a) Under the stated assumptions are both estimators consistent for β?

(b) Are there conditions under which either estimator is efficient?

Exercise 7.10 In the homoskedastic regression model $Y = X'\beta + e$ with $\mathbb{E}[e \mid x] = 0$ and $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$ suppose $\widehat{\beta}$ is the OLS estimator of β with covariance matrix estimator $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ based on a sample of size n. Let $\widehat{\sigma}^2$ be the estimator of σ^2. You wish to forecast an out-of-sample value of Y_{n+1} given that $X_{n+1} = x$. Thus the available information is the sample, the estimates $(\widehat{\beta}, \widehat{\boldsymbol{V}}_{\widehat{\beta}}, \widehat{\sigma}^2)$, the residuals $\widehat{e}_i$, and the out-of-sample value of the regressors X_{n+1}.

(a) Find a point forecast of Y_{n+1}.

(b) Find an estimator of the variance of this forecast.

Exercise 7.11 Take a regression model with i.i.d. observations (Y_i, X_i) with $X \in \mathbb{R}$

$$Y = X\beta + e$$
$$\mathbb{E}[e \mid X] = 0$$
$$\Omega = \mathbb{E}\left[X^2 e^2\right].$$

Let $\widehat{\beta}$ be the OLS estimator of β with residuals $\widehat{e}_i = Y_i - X_i\widehat{\beta}$. Consider the estimators of Ω

$$\widetilde{\Omega} = \frac{1}{n}\sum_{i=1}^n X_i^2 e_i^2$$
$$\widehat{\Omega} = \frac{1}{n}\sum_{i=1}^n X_i^2 \widehat{e}_i^2.$$

(a) Find the asymptotic distribution of $\sqrt{n}\left(\widetilde{\Omega} - \Omega\right)$ as $n \to \infty$.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\Omega} - \Omega\right)$ as $n \to \infty$.

(c) How do you use the regression assumption $\mathbb{E}[e_i \mid X_i] = 0$ in your answer to (b)?

Exercise 7.12 Consider the model

$$Y = \alpha + \beta X + e$$
$$\mathbb{E}[e] = 0$$
$$\mathbb{E}[Xe] = 0$$

with both Y and X scalar. Assuming $\alpha > 0$ and $\beta < 0$ suppose the parameter of interest is the area under the regression curve (e.g. consumer surplus), which is $A = -\alpha^2/2\beta$.

Let $\widehat{\theta} = (\widehat{\alpha}, \widehat{\beta})'$ be the least squares estimators of $\theta = (\alpha, \beta)'$ so that $\sqrt{n}\left(\widehat{\theta} - \theta\right) \to_d N(0, \boldsymbol{V}_\theta)$ and let $\widehat{\boldsymbol{V}}_\theta$ be a standard estimator for $\boldsymbol{V}_\theta$.

(a) Given the above, describe an estimator of A.

(b) Construct an asymptotic $1-\eta$ confidence interval for A.

Exercise 7.13 Consider an i.i.d. sample $\{Y_i, X_i\}$ $i = 1, ..., n$ where Y and X are scalar. Consider the reverse projection model $X = Y\gamma + u$ with $\mathbb{E}[Yu] = 0$ and define the parameter of interest as $\theta = 1/\gamma$.

(a) Propose an estimator $\widehat{\gamma}$ of γ.

(b) Propose an estimator $\widehat{\theta}$ of θ.

(c) Find the asymptotic distribution of $\widehat{\theta}$.

(d) Find an asymptotic standard error for $\widehat{\theta}$.

Exercise 7.14 Take the model

$$\begin{aligned} Y &= X_1\beta_1 + X_2\beta_2 + e \\ \mathbb{E}[Xe] &= 0 \end{aligned}$$

with both $\beta_1 \in \mathbb{R}$ and $\beta_2 \in \mathbb{R}$, and define the parameter $\theta = \beta_1\beta_2$.

(a) What is the appropriate estimator $\widehat{\theta}$ for θ?

(b) Find the asymptotic distribution of $\widehat{\theta}$ under standard regularity conditions.

(c) Show how to calculate an asymptotic 95% confidence interval for θ.

Exercise 7.15 Take the linear model $Y = X\beta + e$ with $\mathbb{E}[e \mid X] = 0$ and $X \in \mathbb{R}$. Consider the estimator

$$\widehat{\beta} = \frac{\sum_{i=1}^n X_i^3 Y_i}{\sum_{i=1}^n X_i^4}.$$

Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ as $n \to \infty$.

Exercise 7.16 From an i.i.d. sample (Y_i, X_i) of size n you randomly take half the observations. You estimate a least squares regression of Y on X using only this sub-sample. Is the estimated slope coefficient $\widehat{\beta}$ consistent for the population projection coefficient? Explain your reasoning.

Exercise 7.17 An economist reports a set of parameter estimates, including the coefficient estimates $\widehat{\beta}_1 = 1.0$, $\widehat{\beta}_2 = 0.8$, and standard errors $s(\widehat{\beta}_1) = 0.07$ and $s(\widehat{\beta}_2) = 0.07$. The author writes "The estimates show that β_1 is larger than β_2."

(a) Write down the formula for an asymptotic 95% confidence interval for $\theta = \beta_1 - \beta_2$, expressed as a function of $\widehat{\beta}_1$, $\widehat{\beta}_2$, $s(\widehat{\beta}_1)$, $s(\widehat{\beta}_2)$ and $\widehat{\rho}$, where $\widehat{\rho}$ is the estimated correlation between $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

(b) Can $\widehat{\rho}$ be calculated from the reported information?

(c) Is the author correct? Does the reported information support the author's claim?

Exercise 7.18 Suppose an economic model suggests

$$m(x) = \mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x + \beta_2 x^2$$

where $X \in \mathbb{R}$. You have a random sample (Y_i, X_i), $i = 1, ..., n$.

(a) Describe how to estimate $m(x)$ at a given value x.

(b) Describe (be specific) an appropriate confidence interval for $m(x)$.

Exercise 7.19 Take the model $Y = X'\beta + e$ with $\mathbb{E}[Xe] = 0$ and suppose you have observations $i = 1, ..., 2n$. (The number of observations is $2n$.) You randomly split the sample in half, (each has n observations), calculate $\widehat{\beta}_1$ by least squares on the first sample, and $\widehat{\beta}_2$ by least squares on the second sample. What is the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)$?

Exercise 7.20 The variables $\{Y_i, X_i, W_i\}$ are a random sample. The parameter β is estimated by minimizing the criterion function

$$S(\beta) = \sum_{i=1}^{n} W_i \left(Y_i - X_i'\beta\right)^2$$

That is $\widehat{\beta} = \operatorname{argmin}_\beta S(\beta)$.

(a) Find an explicit expression for $\widehat{\beta}$.

(b) What population parameter β is $\widehat{\beta}$ estimating? Be explicit about any assumptions you need to impose. Do not make more assumptions than necessary.

(c) Find the probability limit for $\widehat{\beta}$ as $n \to \infty$.

(d) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\beta} - \beta\right)$ as $n \to \infty$.

Exercise 7.21 Take the model

$$\begin{aligned} Y &= X'\beta + e \\ \mathbb{E}[e \mid X] &= 0 \\ \mathbb{E}\left[e^2 \mid X\right] &= Z'\gamma \end{aligned}$$

where Z is a (vector) function of X. The sample is $i = 1, ..., n$ with i.i.d. observations. Assume that $Z'\gamma > 0$ for all Z. Suppose you want to forecast Y_{n+1} given $X_{n+1} = x$ and $Z_{n+1} = z$ for an out-of-sample observation $n+1$. Describe how you would construct a point forecast and a forecast interval for Y_{n+1}.

Exercise 7.22 Take the model

$$\begin{aligned} Y &= X'\beta + e \\ \mathbb{E}[e \mid X] &= 0 \\ Z &= X'\beta\gamma + u \\ \mathbb{E}[u \mid X] &= 0 \end{aligned}$$

where X is a k vector and Z is scalar. Your goal is to estimate the scalar parameter γ. You use a two-step estimator:

- Estimate $\widehat{\beta}$ by least squares of Y on X.
- Estimate $\widehat{\gamma}$ by least squares of Z on $X'\widehat{\beta}$.

(a) Show that $\widehat{\gamma}$ is consistent for γ.

(b) Find the asymptotic distribution of $\widehat{\gamma}$ when $\gamma = 0$.

Exercise 7.23 The model is $Y = X + e$ with $\mathbb{E}[e \mid X] = 0$ and $X \in \mathbb{R}$. Consider the estimator

$$\widetilde{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i}{X_i}.$$

Find conditions under which $\widetilde{\beta}$ is consistent for β as $n \to \infty$.

Exercise 7.24 The parameter β is defined in the model $Y = X^*\beta + e$ where e is independent of $X^* \geq 0$, $\mathbb{E}[e] = 0$, $\mathbb{E}\left[e^2\right] = \sigma^2$. The observables are (Y, X) where $X = X^* v$ and $v > 0$ is random scale measurement error, independent of X^* and e. Consider the least squares estimator $\widehat{\beta}$ for β.

(a) Find the plim of $\widehat{\beta}$ expressed in terms of β and moments of (X, v, e).

(b) Can you find a non-trivial condition under which $\widehat{\beta}$ is consistent for β? (By non-trivial we mean something other than $v = 1$.)

Exercise 7.25 Take the projection model $Y = X'\beta + e$ with $\mathbb{E}[Xe] = 0$. For a positive function $w(x)$ let $W_i = w(X_i)$. Consider the estimator

$$\widetilde{\beta} = \left(\sum_{i=1}^{n} W_i X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} W_i X_i Y_i\right).$$

Find the probability limit (as $n \to \infty$) of $\widetilde{\beta}$. Do you need to add an assumption? Is $\widetilde{\beta}$ consistent for $\widetilde{\beta}$? If not, under what assumption is $\widetilde{\beta}$ consistent for β?

Exercise 7.26 Take the regression model

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

with $X \in \mathbb{R}^k$. Assume that $\mathbb{P}[e = 0] = 0$. Consider the infeasible estimator

$$\widetilde{\beta} = \left(\sum_{i=1}^{n} e_i^{-2} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{n} e_i^{-2} X_i Y_i\right).$$

This is a WLS estimator using the weights e_i^{-2}.

(a) Find the asymptotic distribution of $\widetilde{\beta}$.

(b) Contrast your result with the asymptotic distribution of infeasible GLS.

Exercise 7.27 The model is $Y = X'\beta + e$ with $\mathbb{E}[e \mid X] = 0$. An econometrician is worried about the impact of some unusually large values of the regressors. The model is thus estimated on the subsample for which $|X_i| \leq c$ for some fixed c. Let $\widetilde{\beta}$ denote the OLS estimator on this subsample. It equals

$$\widetilde{\beta} = \left(\sum_{i=1}^{n} X_i X_i' \mathbb{1}\{|X_i| \leq c\}\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i \mathbb{1}\{|X_i| \leq c\}\right).$$

(a) Show that $\widetilde{\beta} \underset{p}{\longrightarrow} \beta$.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widetilde{\beta} - \beta\right)$.

Exercise 7.28 As in Exercise 3.26, use the `cps09mar` 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 coefficient estimates 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 for *experience*= 10. 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 $\widehat{\beta}$. Compute $s(\widehat{\theta})$ from the estimated model.

(d) Construct a 90% asymptotic confidence interval for θ from the estimated model.

(e) Compute the regression function at *education*= 12 and *experience*= 20. Compute a 95% confidence interval for the regression function at this point.

(f) Consider an out-of-sample individual with 16 years of education and 5 years experience. Construct an 80% forecast interval for their log wage and wage. [To obtain the forecast interval for the wage, apply the exponential function to both endpoints.]

Chapter 8

Restricted Estimation

8.1 Introduction

In the linear projection model

$$\begin{aligned} Y &= X'\beta + e \\ \mathbb{E}[Xe] &= 0 \end{aligned}$$

a common task is to impose a constraint on the coefficient vector β. For example, partitioning $X' = \left(X_1', X_2'\right)$ and $\beta' = \left(\beta_1', \beta_2'\right)$ a typical constraint is an exclusion restriction of the form $\beta_2 = 0$. In this case the constrained model is

$$\begin{aligned} Y &= X_1'\beta_1 + e \\ \mathbb{E}[Xe] &= 0. \end{aligned}$$

At first glance this appears the same as the linear projection model but there is one important difference: the error e is uncorrelated with the entire regressor vector $X' = \left(X_1', X_2'\right)$ not just the included regressor X_1.

In general, a set of q linear constraints on β takes the form

$$\boldsymbol{R}'\beta = \boldsymbol{c} \tag{8.1}$$

where $\boldsymbol{R}$ is $k \times q$, $\operatorname{rank}(\boldsymbol{R}) = q < k$, and $\boldsymbol{c}$ is $q \times 1$. The assumption that $\boldsymbol{R}$ is full rank means that the constraints are linearly independent (there are no redundant or contradictory constraints). We define the restricted parameter space B as the set of values of β which satisfy (8.1), that is

$$B = \left\{\beta : \boldsymbol{R}'\beta = \boldsymbol{c}\right\}.$$

Sometimes we will call (8.1) a **constraint** and sometimes a **restriction**. They are the same thing. Similarly sometimes we will call estimators which satisfy (8.1) **constrained estimators** and sometimes **restricted estimators**. They mean the same thing.

The constraint $\beta_2 = 0$ discussed above is a special case of the constraint (8.1) with

$$\boldsymbol{R} = \begin{pmatrix} 0 \\ \boldsymbol{I}_{k_2} \end{pmatrix}, \tag{8.2}$$

a selector matrix, and $\boldsymbol{c} = 0$.

Another common restriction is that a set of coefficients sum to a known constant, i.e. $\beta_1 + \beta_2 = 1$. For example, 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 β imposing the linear restriction (8.1)? If we impose such constraints what is the sampling distribution of the resulting estimator? How should we calculate standard errors? These are the questions explored in this chapter.

8.2 Constrained Least Squares

An intuitively appealing method to estimate a constrained linear projection is to minimize the least squares criterion subject to the constraint $\boldsymbol{R}'\beta = \boldsymbol{c}$.

The constrained least squares estimator is

$$\widetilde{\beta}_{\text{cls}} = \underset{\boldsymbol{R}'\beta = \boldsymbol{c}}{\operatorname{argmin}} \operatorname{SSE}(\beta) \tag{8.3}$$

where

$$\operatorname{SSE}(\beta) = \sum_{i=1}^{n} \left(Y_i - X_i'\beta\right)^2 = \boldsymbol{Y}'\boldsymbol{Y} - 2\boldsymbol{Y}'\boldsymbol{X}\beta + \beta'\boldsymbol{X}'\boldsymbol{X}\beta. \tag{8.4}$$

The estimator $\widetilde{\beta}_{\text{cls}}$ minimizes the sum of squared errors over all $\beta \in B$, or equivalently such that the restriction (8.1) holds. We call $\widetilde{\beta}_{\text{cls}}$ the **constrained least squares** (CLS) estimator. We use the convention of using a tilde "~" rather than a hat "^" to indicate that $\widetilde{\beta}_{\text{cls}}$ is a restricted estimator in contrast to the unrestricted least squares estimator $\widehat{\beta}$ and write it as $\widetilde{\beta}_{\text{cls}}$ to be clear that the estimation method is CLS.

One method to find the solution to (8.3) is the technique of Lagrange multipliers. The problem (8.3) is equivalent to finding the critical points of the Lagrangian

$$\mathscr{L}(\beta, \lambda) = \frac{1}{2}\operatorname{SSE}(\beta) + \lambda'\left(\boldsymbol{R}'\beta - \boldsymbol{c}\right) \tag{8.5}$$

over (β, λ) where λ is an $s \times 1$ vector of Lagrange multipliers. The solution is a saddlepoint. The Lagrangian is minimized over β while maximized over λ. The first-order conditions for the solution of (8.5) are

$$\frac{\partial}{\partial \beta}\mathscr{L}(\widetilde{\beta}_{\text{cls}}, \widetilde{\lambda}_{\text{cls}}) = -\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{X}'\boldsymbol{X}\widetilde{\beta}_{\text{cls}} + \boldsymbol{R}\widetilde{\lambda}_{\text{cls}} = 0 \tag{8.6}$$

and

$$\frac{\partial}{\partial \lambda}\mathscr{L}(\widetilde{\beta}_{\text{cls}}, \widetilde{\lambda}_{\text{cls}}) = \boldsymbol{R}'\widetilde{\beta} - \boldsymbol{c} = 0. \tag{8.7}$$

Premultiplying (8.6) by $\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ we obtain

$$-\boldsymbol{R}'\widehat{\beta} + \boldsymbol{R}'\widetilde{\beta}_{\text{cls}} + \boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\widetilde{\lambda}_{\text{cls}} = 0$$

where $\widehat{\beta} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}$ is the unrestricted least squares estimator. Imposing $\boldsymbol{R}'\widetilde{\beta}_{\text{cls}} - \boldsymbol{c} = 0$ from (8.7) and solving for $\widetilde{\lambda}_{\text{cls}}$ we find

$$\widetilde{\lambda}_{\text{cls}} = \left[\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c}\right).$$

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.10.)

Substituting this expression into (8.6) and solving for $\widetilde{\beta}_{\text{cls}}$ we find the solution to the constrained minimization problem (8.3)

$$\widetilde{\beta}_{\text{cls}} = \widehat{\beta}_{\text{ols}} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left[\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\beta}_{\text{ols}} - \boldsymbol{c}\right). \tag{8.8}$$

(See Exercise 8.5 to verify that (8.8) satisfies (8.1).)

This is a general formula for the CLS estimator. It also can be written as

$$\widetilde{\beta}_{\text{cls}} = \widehat{\beta}_{\text{ols}} - \widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta}_{\text{ols}} - \boldsymbol{c}\right). \tag{8.9}$$

The CLS residuals are $\widetilde{e}_i = Y_i - X_i'\widetilde{\beta}_{\text{cls}}$ and are written in vector notation as $\widetilde{\boldsymbol{e}}$.

To illustrate we generated a random sample of 100 observations for the variables (Y, X_1, X_2) and calculated the sum of squared errors function for the regression of Y on X_1 and X_2. Figure 8.1 displays contour plots of the sum of squared errors function. The center of the contour plots is the least squares minimizer $\widehat{\beta}_{\text{ols}} = (0.33, 0.26)'$. Suppose it is desired to estimate the coefficients subject to the constraint $\beta_1 + \beta_2 = 1$. This constraint is displayed in the figure by the straight line. The constrained least squares estimator is the point on this straight line which yields the smallest sum of squared errors. This is the point which intersects with the lowest contour plot. The solution is the point where a contour plot is tangent to the constraint line and is marked as $\widetilde{\beta}_{\text{cls}} = (0.52, 0.48)'$.

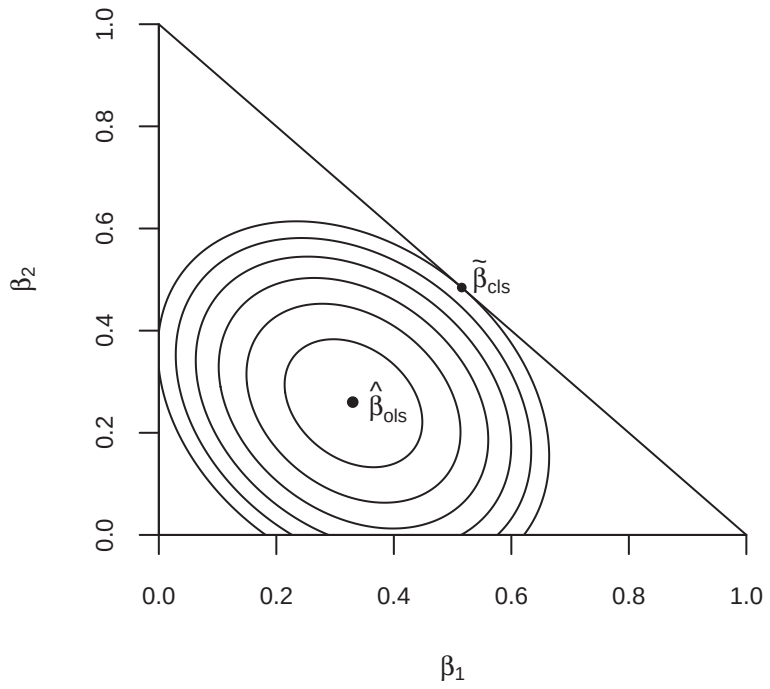


Figure 8.1: Constrained Least Squares Criterion

In Stata constrained least squares is implemented using the `cnsreg` command.

8.3 Exclusion Restriction

While (8.8) is a general formula for CLS, 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 that the unconstrained model is

$$Y = X_1'\beta_1 + X_2'\beta_2 + e, \tag{8.10}$$

the exclusion restriction is $\beta_2 = 0$, and the constrained equation is

$$Y = X_1'\beta_1 + e. \tag{8.11}$$

In this setting the CLS estimator is OLS of Y on X_1. (See Exercise 8.1.) We can write this as

$$\widetilde{\beta}_1 = \left(\sum_{i=1}^n X_{1i}X_{1i}'\right)^{-1}\left(\sum_{i=1}^n X_{1i}Y_i\right). \tag{8.12}$$

The CLS estimator of the entire vector $\beta' = \left(\beta_1', \beta_2'\right)$ is

$$\widetilde{\beta} = \begin{pmatrix} \widetilde{\beta}_1 \\ 0 \end{pmatrix}. \tag{8.13}$$

It is not immediately obvious but (8.8) and (8.13) are algebraically identical. To see this the first component of (8.8) with (8.2) is

$$\widetilde{\beta}_1 = \begin{pmatrix} \boldsymbol{I}_{k_2} & 0 \end{pmatrix}\left[\widehat{\beta} - \widehat{\boldsymbol{Q}}_{XX}^{-1}\begin{pmatrix} 0 \\ \boldsymbol{I}_{k_2} \end{pmatrix}\left[\begin{pmatrix} 0 & \boldsymbol{I}_{k_2} \end{pmatrix}\widehat{\boldsymbol{Q}}_{XX}^{-1}\begin{pmatrix} 0 \\ \boldsymbol{I}_{k_2} \end{pmatrix}\right]^{-1}\begin{pmatrix} 0 & \boldsymbol{I}_{k_2} \end{pmatrix}\widehat{\beta}\right].$$

Using (3.39) this equals

$$\begin{aligned}
\widetilde{\beta}_1 &= \widehat{\beta}_1 - \widehat{\boldsymbol{Q}}^{12}\left(\widehat{\boldsymbol{Q}}^{22}\right)^{-1}\widehat{\beta}_2 \\
&= \widehat{\beta}_1 + \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{22\cdot 1}\widehat{\beta}_2 \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{1Y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{2Y}\right) \\
&+ \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{22\cdot 1}\widehat{\boldsymbol{Q}}_{22\cdot 1}^{-1}\left(\widehat{\boldsymbol{Q}}_{2y} - \widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1Y}\right) \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{1Y} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1Y}\right) \\
&= \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\left(\widehat{\boldsymbol{Q}}_{11} - \widehat{\boldsymbol{Q}}_{12}\widehat{\boldsymbol{Q}}_{22}^{-1}\widehat{\boldsymbol{Q}}_{21}\right)\widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1Y} \\
&= \widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1Y}
\end{aligned}$$

which is (8.13) as originally claimed.

8.4 Finite Sample Properties

In this section we explore some of the properties of the CLS estimator in the linear regression model

$$Y = X'\beta + e \tag{8.14}$$

$$\mathbb{E}[e \mid X] = 0. \tag{8.15}$$

First, it is useful to write the estimator and the residuals as linear functions of the error vector. These are algebraic relationships and do not rely on the linear regression assumptions.

Theorem 8.1 The CLS estimator satisfies

1. $\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c} = \boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{e}$
2. $\widetilde{\beta}_{\text{cls}} - \beta = \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' - \boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}$
3. $\widetilde{\boldsymbol{e}} = \left(\boldsymbol{I} - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}$
4. $\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'$ is symmetric and idempotent
5. $\operatorname{tr}\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right) = n - k + q$

where $\boldsymbol{P} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'$ and $\boldsymbol{A} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$.

For a proof see Exercise 8.6.

Given the linearity of Theorem 8.1.2 it is not hard to show that the CLS estimator is unbiased for β.

Theorem 8.2 In the linear regression model (8.14)-(8.15) under (8.1), $\mathbb{E}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right] = \beta$.

For a proof see Exercise 8.7.

We can also calculate the covariance matrix of $\widetilde{\beta}_{\text{cls}}$. First, for simplicity take the case of conditional homoskedasticity.

Theorem 8.3 In the homoskedastic linear regression model (8.14)-(8.15) with $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$, under (8.1),

$$\begin{aligned}\boldsymbol{V}^0_{\widetilde{\beta}} &= \operatorname{var}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right] \\ &= \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)\sigma^2.\end{aligned}$$

For a proof see Exercise 8.8.

We use the $\boldsymbol{V}^0_{\widetilde{\beta}}$ notation to emphasize that this is the covariance matrix under the assumption of conditional homoskedasticity.

For inference we need an estimate of $\boldsymbol{V}^0_{\widetilde{\beta}}$. A natural estimator is

$$\widehat{\boldsymbol{V}}^0_{\widetilde{\beta}} = \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)s^2_{\text{cls}}$$

where

$$s^2_{\text{cls}} = \frac{1}{n-k+q}\sum_{i=1}^{n}\widetilde{e}_i^2 \tag{8.16}$$

is a biased-corrected estimator of σ^2. Standard errors for the components of β are then found by taking the squares roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widetilde{\beta}}$, for example

$$s(\widehat{\beta}_j) = \sqrt{\left[\widehat{\boldsymbol{V}}^0_{\widetilde{\beta}}\right]_{jj}}.$$

The estimator (8.16) has the property that it is unbiased for σ^2 under conditional homoskedasticity. To see this, using the properties of Theorem 8.1,

$$\begin{aligned}
\left(n-k+q\right)s^2_{\text{cls}} &= \widetilde{\boldsymbol{e}}'\widetilde{\boldsymbol{e}} \\
&= \boldsymbol{e}'\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e} \\
&= \boldsymbol{e}'\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}. \qquad (8.17)
\end{aligned}$$

We defer the remainder of the proof to Exercise 8.9.

Theorem 8.4 In the homoskedastic linear regression model (8.14)-(8.15) with $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2$, under (8.1), $\mathbb{E}\left[s^2_{\text{cls}} \mid \boldsymbol{X}\right] = \sigma^2$ and $\mathbb{E}\left[\widehat{\boldsymbol{V}}^0_{\widetilde{\beta}} \mid \boldsymbol{X}\right] = \boldsymbol{V}^0_{\widetilde{\beta}}$.

Now consider the distributional properties in the normal regression model $Y = X'\beta + e$ with $e \sim \text{N}(0, \sigma^2)$. By the linearity of Theorem 8.1.2, conditional on $\boldsymbol{X}$, $\widetilde{\beta}_{\text{cls}} - \beta$ is normal. Given Theorems 8.2 and 8.3 we deduce that $\widetilde{\beta}_{\text{cls}} \sim \text{N}\left(\beta, \boldsymbol{V}^0_{\widetilde{\beta}}\right)$.

Similarly, from Exericise 8.1 we know $\widetilde{\boldsymbol{e}} = \left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\boldsymbol{e}$ is linear in $\boldsymbol{e}$ so is also conditionally normal. Furthermore, since $\left(\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'\right)\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} - \boldsymbol{X}\boldsymbol{A}\right) = 0$, $\widetilde{\boldsymbol{e}}$ and $\widetilde{\beta}_{\text{cls}}$ are uncorrelated and thus independent. Thus s^2_{cls} and $\widetilde{\beta}_{\text{cls}}$ are independent.

From (8.17) and the fact that $\boldsymbol{I}_n - \boldsymbol{P} + \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}'$ is idempotent with rank $n-k+q$ it follows that

$$s^2_{\text{cls}} \sim \sigma^2 \chi^2_{n-k+q} / \left(n-k+q\right).$$

It follows that the t-statistic has the exact distribution

$$T = \frac{\widehat{\beta}_j - \beta_j}{s(\widehat{\beta}_j)} \sim \frac{\text{N}(0,1)}{\sqrt{\chi^2_{n-k+q} \Big/ (n-k+q)}} \sim t_{n-k+q}$$

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

The relevance of this calculation is that the "degrees of freedom" for CLS regression equal $n-k+q$ rather than $n-k$ as in OLS. Essentially the model has $k-q$ free parameters instead of k. Another way of thinking about this is that estimation of a model with k coefficients and q restrictions is equivalent to estimation with $k-q$ coefficients.

We summarize the properties of the normal regression model.

Theorem 8.5 In the normal linear regression model (8.14)-(8.15) with constraint (8.1),

$$\begin{aligned}\widetilde{\beta}_{\text{cls}} &\sim \mathrm{N}\left(\beta, \boldsymbol{V}_{\widetilde{\beta}}^{0}\right)\\ \frac{(n-k+q)\, s_{\text{cls}}^{2}}{\sigma^{2}} &\sim \chi_{n-k+q}^{2}\\ T &\sim t_{n-k+q}.\end{aligned}$$

An interesting relationship is that in the homoskedastic regression model

$$\begin{aligned}\operatorname{cov}\left(\widehat{\beta}_{\text{ols}}-\widetilde{\beta}_{\text{cls}}, \widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right) &= \mathbb{E}\left[\left(\widehat{\beta}_{\text{ols}}-\widetilde{\beta}_{\text{cls}}\right)\left(\widetilde{\beta}_{\text{cls}}-\beta\right)' \mid \boldsymbol{X}\right]\\ &= \mathbb{E}\left[\boldsymbol{A}\boldsymbol{X}'\boldsymbol{e}\boldsymbol{e}'\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}-\boldsymbol{X}\boldsymbol{A}\right) \mid \boldsymbol{X}\right]\\ &= \boldsymbol{A}\boldsymbol{X}'\left(\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}-\boldsymbol{X}\boldsymbol{A}\right)\sigma^{2}=0.\end{aligned}$$

This means that $\widehat{\beta}_{\text{ols}}-\widetilde{\beta}_{\text{cls}}$ and $\widetilde{\beta}_{\text{cls}}$ are conditionally uncorrelated and hence independent. A corollary is

$$\operatorname{cov}\left(\widehat{\beta}_{\text{ols}}, \widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right) = \operatorname{var}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right].$$

A second corollary is

$$\begin{aligned}\operatorname{var}\left[\widehat{\beta}_{\text{ols}}-\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right] &= \operatorname{var}\left[\widehat{\beta}_{\text{ols}} \mid \boldsymbol{X}\right]-\operatorname{var}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right]\\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^{2}.\end{aligned} \tag{8.18}$$

This also shows that the difference between the CLS and OLS variances matrices equals

$$\operatorname{var}\left[\widehat{\beta}_{\text{ols}} \mid \boldsymbol{X}\right]-\operatorname{var}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\sigma^{2} \geq 0$$

the final equality meaning positive semi-definite. It follows that $\operatorname{var}\left[\widehat{\beta}_{\text{ols}} \mid \boldsymbol{X}\right] \geq \operatorname{var}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right]$ in the positive definite sense, and thus CLS is more efficient than OLS. Both estimators are unbiased (in the linear regression model) and CLS has a lower covariance matrix (in the linear homoskedastic regression model).

The relationship (8.18) is rather interesting and will appear again. The expression says that the variance of the difference between the estimators is equal to the difference between the variances. This is rather special. It occurs generically when we are comparing an efficient and an inefficient estimator. We call (8.18) the **Hausman Equality** as it was first pointed out in econometrics by Hausman (1978).

8.5 Minimum Distance

The previous section explored the finite sample distribution theory under the assumptions of the linear regression model, homoskedastic regression model, and normal regression model. We now return to the general projection model where we do not impose linearity, homoskedasticity, nor normality. We are interested in the question: Can we do better than CLS in this setting?

A minimum distance estimator tries to find a parameter value satisfying the constraint which is as close as possible to the unconstrained estimator. Let $\widehat{\beta}$ be the unconstrained least squares estimator, and for some $k \times k$ positive definite weight matrix $\widehat{\boldsymbol{W}}$ define the quadratic criterion function

$$J\left(\beta\right) = n\left(\widehat{\beta}-\beta\right)'\widehat{\boldsymbol{W}}\left(\widehat{\beta}-\beta\right). \tag{8.19}$$

This is a (squared) weighted Euclidean distance between $\widehat{\beta}$ and β. $J(\beta)$ is small if β is close to $\widehat{\beta}$, and is minimized at zero only if $\beta = \widehat{\beta}$. A **minimum distance estimator** $\widetilde{\beta}_{\text{md}}$ for β minimizes $J(\beta)$ subject to the constraint (8.1), that is,

$$\widetilde{\beta}_{\text{md}} = \underset{\boldsymbol{R}'\beta=\boldsymbol{c}}{\operatorname{argmin}}\ J(\beta).$$

The CLS estimator is the special case when $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{XX}$ and we write this criterion function as

$$J^0(\beta) = n(\widehat{\beta}-\beta)'\widehat{\boldsymbol{Q}}_{XX}(\widehat{\beta}-\beta). \tag{8.20}$$

To see the equality of CLS and minimum distance rewrite the least squares criterion as follows. Substitute the unconstrained least squares fitted equation $Y_i = X_i'\widehat{\beta} + \widehat{e}_i$ into SSE(β) to obtain

$$\begin{aligned}
\text{SSE}(\beta) &= \sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2 \\
&= \sum_{i=1}^{n}\left(X_i'\widehat{\beta} + \widehat{e}_i - X_i'\beta\right)^2 \\
&= \sum_{i=1}^{n}\widehat{e}_i^2 + (\widehat{\beta}-\beta)'\left(\sum_{i=1}^{n}X_iX_i'\right)(\widehat{\beta}-\beta) \\
&= n\widehat{\sigma}^2 + J^0(\beta)
\end{aligned} \tag{8.21}$$

where the third equality uses the fact that $\sum_{i=1}^n X_i\widehat{e}_i = 0$, and the last line uses $\sum_{i=1}^n X_iX_i' = n\widehat{\boldsymbol{Q}}_{XX}$. The expression (8.21) only depends on β through $J^0(\beta)$. Thus minimization of SSE(β) and $J^0(\beta)$ are equivalent, and hence $\widetilde{\beta}_{\text{md}} = \widetilde{\beta}_{\text{cls}}$ when $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{XX}$.

We can solve for $\widetilde{\beta}_{\text{md}}$ explicitly by the method of Lagrange multipliers. The Lagrangian is

$$\mathscr{L}(\beta,\lambda) = \frac{1}{2}J(\beta,\widehat{\boldsymbol{W}}) + \lambda'(\boldsymbol{R}'\beta - \boldsymbol{c}).$$

The solution to the pair of first order conditions is

$$\widetilde{\lambda}_{\text{md}} = n\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right) \tag{8.22}$$

$$\widetilde{\beta}_{\text{md}} = \widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c}\right). \tag{8.23}$$

(See Exercise 8.10.) Comparing (8.23) with (8.9) we can see that $\widetilde{\beta}_{\text{md}}$ specializes to $\widetilde{\beta}_{\text{cls}}$ when we set $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{XX}$.

An obvious question is which weight matrix $\widehat{\boldsymbol{W}}$ is best. We will address this question after we derive the asymptotic distribution for a general weight matrix.

8.6 Asymptotic Distribution

We first show that the class of minimum distance estimators are consistent for the population parameters when the constraints are valid.

Assumption 8.1 $\boldsymbol{R}'\beta = \boldsymbol{c}$ where $\boldsymbol{R}$ is $k \times q$ with $\operatorname{rank}(\boldsymbol{R}) = q$.

Assumption 8.2 $\widehat{\boldsymbol{W}} \xrightarrow[p]{} \boldsymbol{W} > 0$.

Theorem 8.6 Consistency
Under Assumptions 7.1, 8.1, and 8.2, $\widetilde{\beta}_{\text{md}} \xrightarrow[p]{} \beta$ as $n \to \infty$.

For a proof see Exercise 8.11.

Theorem 8.6 shows that consistency holds for any weight matrix with a positive definite limit so includes the CLS estimator.

Similarly, the constrained estimators are asymptotically normally distributed.

Theorem 8.7 Asymptotic Normality
Under Assumptions 7.2, 8.1, and 8.2,

$$\sqrt{n}\left(\widetilde{\beta}_{\text{md}} - \beta\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_\beta(\boldsymbol{W})\right)$$

as $n \to \infty$, where

$$\begin{aligned} \boldsymbol{V}_\beta(\boldsymbol{W}) = \boldsymbol{V}_\beta &- \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_\beta \\ &- \boldsymbol{V}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}^{-1} \\ &+ \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{W}^{-1} \end{aligned} \tag{8.24}$$

and $\boldsymbol{V}_\beta = \boldsymbol{Q}_{XX}^{-1}\Omega\boldsymbol{Q}_{XX}^{-1}$.

For a proof see Exercise 8.12.

Theorem 8.7 shows that the minimum distance estimator is asymptotically normal for all positive definite weight matrices. The asymptotic variance depends on $\boldsymbol{W}$. The theorem includes the CLS estimator as a special case by setting $\boldsymbol{W} = \boldsymbol{Q}_{XX}$.

Theorem 8.8 Asymptotic Distribution of CLS Estimator
Under Assumptions 7.2 and 8.1, as $n \to \infty$

$$\sqrt{n}\left(\widetilde{\beta}_{\text{cls}} - \beta\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_{\text{cls}}\right)$$

where

$$\begin{aligned} \boldsymbol{V}_{\text{cls}} = \boldsymbol{V}_\beta &- \boldsymbol{Q}_{XX}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_\beta \\ &- \boldsymbol{V}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}_{XX}^{-1} \\ &+ \boldsymbol{Q}_{XX}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{Q}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{Q}_{XX}^{-1}. \end{aligned}$$

For a proof see Exercise 8.13.

8.7 Variance Estimation and Standard Errors

Earlier we introduced the covariance matrix estimator under the assumption of conditional homoskedasticity. We now introduce an estimator which does not impose homoskedasticity.

The asymptotic covariance matrix $\boldsymbol{V}_{\text{cls}}$ may be estimated by replacing $\boldsymbol{V}_\beta$ with a consistent estimator such as $\widehat{\boldsymbol{V}}_\beta$. A more efficient estimator can be obtained by using the restricted coefficient estimator which we now show. Given the constrained least squares squares residuals $\widetilde{e}_i = Y_i - X_i'\widetilde{\beta}_{\text{cls}}$ we can estimate the matrix $\Omega = \mathbb{E}\left[XX'e^2\right]$ by

$$\widetilde{\Omega} = \frac{1}{n-k+q}\sum_{i=1}^{n} X_i X_i' \widetilde{e}_i^2.$$

Notice that we have used an adjusted degrees of freedom. This is an *ad hoc* adjustment designed to mimic that used for estimation of the error variance σ^2. The moment estimator of $\boldsymbol{V}_\beta$ is

$$\widetilde{\boldsymbol{V}}_\beta = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widetilde{\Omega}\widehat{\boldsymbol{Q}}_{XX}^{-1}$$

and that for $\boldsymbol{V}_{\text{cls}}$ is

$$\begin{aligned}\widetilde{\boldsymbol{V}}_{\text{cls}} = \widetilde{\boldsymbol{V}}_\beta &- \widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widetilde{\boldsymbol{V}}_\beta \\ &- \widetilde{\boldsymbol{V}}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{\boldsymbol{xx}}^{-1} \\ &+ \widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widetilde{\boldsymbol{V}}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{XX}^{-1}.\end{aligned}$$

We can calculate standard errors for any linear combination $h'\widetilde{\beta}_{\text{cls}}$ such that h does not lie in the range space of $\boldsymbol{R}$. A standard error for $h'\widetilde{\beta}$ is

$$s\left(h'\widetilde{\boldsymbol{\beta}}_{\text{cls}}\right) = \left(n^{-1}h'\widetilde{\boldsymbol{V}}_{\text{cls}}h\right)^{1/2}.$$

8.8 Efficient Minimum Distance Estimator

Theorem 8.7 shows that 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 $\boldsymbol{V}_\beta(\boldsymbol{W})$. This turns out to be $\boldsymbol{W} = \boldsymbol{V}_\beta^{-1}$ as is shown in Theorem 8.9 below. Since $\boldsymbol{V}_\beta^{-1}$ is unknown this weight matrix cannot be used for a feasible estimator but we can replace $\boldsymbol{V}_\beta^{-1}$ with a consistent estimator $\widehat{\boldsymbol{V}}_\beta^{-1}$ and the asymptotic distribution (and efficiency) are unchanged. We call the minimum distance estimator with $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{V}}_\beta^{-1}$ the **efficient minimum distance estimator** and takes the form

$$\widetilde{\beta}_{\text{emd}} = \widehat{\beta} - \widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c}\right). \tag{8.25}$$

The asymptotic distribution of (8.25) can be deduced from Theorem 8.7. (See Exercises 8.14 and 8.15, and the proof in Section 8.16.)

Theorem 8.9 Efficient Minimum Distance Estimator
Under Assumptions 7.2 and 8.1,

$$\sqrt{n}\left(\widetilde{\beta}_{\text{emd}} - \beta\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_{\beta,\text{emd}}\right)$$

as $n \to \infty$, where

$$\boldsymbol{V}_{\beta,\text{emd}} = \boldsymbol{V}_\beta - \boldsymbol{V}_\beta \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_\beta \boldsymbol{R}\right)^{-1} \boldsymbol{R}'\boldsymbol{V}_\beta. \tag{8.26}$$

Since

$$\boldsymbol{V}_{\beta,\text{emd}} \leq \boldsymbol{V}_\beta \tag{8.27}$$

the estimator (8.25) has lower asymptotic variance than the unrestricted estimator. Furthermore, for any $\boldsymbol{W}$,

$$\boldsymbol{V}_{\beta,\text{emd}} \leq \boldsymbol{V}_\beta(\boldsymbol{W}) \tag{8.28}$$

so (8.25) is asymptotically efficient in the class of minimum distance estimators.

Theorem 8.9 shows that the minimum distance estimator with the smallest asymptotic variance is (8.25). 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 $\boldsymbol{W} = \left(\boldsymbol{V}_\beta^0\right)^{-1}$ 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. Standard intuition suggests to apply the same estimation method (least squares) to the unconstrained and constrained models and this is the common empirical practice. But Theorem 8.9 shows that this is inefficient. Why? The reason is that the least squares estimator does not make use of the regressor X_2. It ignores the information $\mathbb{E}[X_2 e] = 0$. This information is relevant when the error is heteroskedastic and the excluded regressors are correlated with the included regressors.

Inequality (8.27) shows that the efficient minimum distance estimator $\widetilde{\beta}_{\text{emd}}$ has a smaller asymptotic variance than the unrestricted least squares estimator $\widehat{\beta}$. This means that efficient estimation is attained by imposing correct restrictions when we use the minimum distance method.

8.9 Exclusion Restriction Revisited

We return to the example of estimation with a simple exclusion restriction. The model is

$$Y = X_1'\beta_1 + X_2'\beta_2 + e$$

with the exclusion restriction $\beta_2 = 0$. We have introduced three estimators of β_1. The first is unconstrained least squares applied to (8.10) which can be written as $\widehat{\beta}_1 = \widehat{\boldsymbol{Q}}_{11\cdot 2}^{-1}\widehat{\boldsymbol{Q}}_{1Y\cdot 2}$. From Theorem 7.25 and equation (7.14) its asymptotic variance is

$$\operatorname{avar}\left[\widehat{\beta}_1\right] = \boldsymbol{Q}_{11\cdot 2}^{-1}\left(\Omega_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\Omega_{21} - \Omega_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} + \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\Omega_{22}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)\boldsymbol{Q}_{11\cdot 2}^{-1}.$$

The second estimator of β_1 is CLS, which can be written as $\widetilde{\beta}_1 = \widehat{\boldsymbol{Q}}_{11}^{-1}\widehat{\boldsymbol{Q}}_{1Y}$. Its asymptotic variance can be deduced from Theorem 8.8, but it is simpler to apply the CLT directly to show that

$$\operatorname{avar}\left[\widetilde{\beta}_1\right] = \boldsymbol{Q}_{11}^{-1}\Omega_{11}\boldsymbol{Q}_{11}^{-1}. \tag{8.29}$$

The third estimator of β_1 is efficient minimum distance. Applying (8.25), it equals

$$\overline{\beta}_1 = \widehat{\beta}_1 - \widehat{\boldsymbol{V}}_{12}\widehat{\boldsymbol{V}}_{22}^{-1}\widehat{\beta}_2 \tag{8.30}$$

where we have partitioned

$$\widehat{\boldsymbol{V}}_{\beta} = \begin{bmatrix} \widehat{\boldsymbol{V}}_{11} & \widehat{\boldsymbol{V}}_{12} \\ \widehat{\boldsymbol{V}}_{21} & \widehat{\boldsymbol{V}}_{22} \end{bmatrix}.$$

From Theorem 8.9 its asymptotic variance is

$$\operatorname{avar}\left[\overline{\beta}_1\right] = \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\boldsymbol{V}_{22}^{-1}\boldsymbol{V}_{21}. \tag{8.31}$$

See Exercise 8.16 to verify equations (8.29), (8.30), and (8.31).

In general the three estimators are different and they have different asymptotic variances. It is instructive to compare the variances to assess whether or not the constrained estimator is more efficient than the unconstrained estimator.

First, assume conditional homoskedasticity. In this case the two covariance matrices simplify to $\operatorname{avar}\left[\widehat{\beta}_1\right] = \sigma^2\boldsymbol{Q}_{11\cdot 2}^{-1}$ and $\operatorname{avar}\left[\widetilde{\beta}_1\right] = \sigma^2\boldsymbol{Q}_{11}^{-1}$. If $\boldsymbol{Q}_{12} = 0$ (so X_1 and X_2 are uncorrelated) then these two variance matrices are equal and the two estimators have equal asymptotic efficiency. Otherwise, since $\boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21} \geq 0$, then $\boldsymbol{Q}_{11} \geq \boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}$ and consequently

$$\boldsymbol{Q}_{11}^{-1}\sigma^2 \leq \left(\boldsymbol{Q}_{11} - \boldsymbol{Q}_{12}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\right)^{-1}\sigma^2.$$

This means that under conditional homoskedasticity $\widetilde{\beta}_1$ has a lower asymptotic covariance matrix than $\widehat{\beta}_1$. Therefore in this context constrained least squares is more efficient than unconstrained least squares. This is consistent with our intuition that imposing a correct restriction (excluding an irrelevant regressor) improves estimation efficiency.

However, in the general case of conditional heteroskedasticity this ranking is not guaranteed. In fact what is really amazing is that the variance ranking can be reversed. The CLS estimator can have a larger asymptotic variance than the unconstrained least squares estimator.

To see this let's use the simple heteroskedastic example from Section 7.4. In that example, $Q_{11} = Q_{22} = 1$, $Q_{12} = \dfrac{1}{2}$, $\Omega_{11} = \Omega_{22} = 1$, and $\Omega_{12} = \dfrac{7}{8}$. We can calculate (see Exercise 8.17) that $Q_{11\cdot 2} = \dfrac{3}{4}$ and

$$\operatorname{avar}\left[\widehat{\beta}_1\right] = \frac{2}{3} \tag{8.32}$$

$$\operatorname{avar}\left[\widetilde{\beta}_1\right] = 1 \tag{8.33}$$

$$\operatorname{avar}\left[\overline{\beta}_1\right] = \frac{5}{8}. \tag{8.34}$$

Thus the CLS estimator $\widetilde{\beta}_1$ has a larger variance than the unrestricted least squares estimator $\widehat{\beta}_1$! 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_2 can actually increase estimation variance, or equivalently, adding an irrelevant variable can decrease the estimation variance.

To repeat this unexpected finding, we have shown that it is possible for least squares applied to the short regression (8.11) to be less efficient for estimation of β_1 than least squares applied to the long regression (8.10) even though the constraint $\beta_2 = 0$ 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 necessarily CLS. While least squares is asymptotically efficient for estimation of the unconstrained projection model it is not an efficient estimator of the constrained projection model.

8.10 Variance and Standard Error Estimation

We have discussed covariance matrix estimation for CLS but not yet for the EMD estimator.

The asymptotic covariance matrix (8.26) may be estimated by replacing $\boldsymbol{V}_\beta$ with a consistent estimator. It is best to construct the variance estimate using $\widetilde{\beta}_{\text{emd}}$. The EMD residuals are $\widetilde{e}_i = Y_i - X_i'\widetilde{\beta}_{\text{emd}}$. Using these we can estimate the matrix $\Omega = \mathbb{E}\left[XX'e^2\right]$ by

$$\widetilde{\Omega} = \frac{1}{n-k+q}\sum_{i=1}^{n} X_i X_i' \widetilde{e}_i^2.$$

Following the formula for CLS we recommend an adjusted degrees of freedom. Given $\widetilde{\Omega}$ the moment estimator of $\boldsymbol{V}_\beta$ is $\widetilde{\boldsymbol{V}}_\beta = \widehat{\boldsymbol{Q}}_{XX}^{-1}\widetilde{\Omega}\widehat{\boldsymbol{Q}}_{XX}^{-1}$. Given this, we construct the variance estimator

$$\widetilde{\boldsymbol{V}}_{\beta,\text{emd}} = \widetilde{\boldsymbol{V}}_\beta - \widetilde{\boldsymbol{V}}_\beta \boldsymbol{R}\left(\boldsymbol{R}'\widetilde{\boldsymbol{V}}_\beta \boldsymbol{R}\right)^{-1}\boldsymbol{R}'\widetilde{\boldsymbol{V}}_\beta. \tag{8.35}$$

A standard error for $h'\widetilde{\beta}$ is then

$$s\left(h'\widetilde{\beta}\right) = \left(n^{-1}h'\widetilde{\boldsymbol{V}}_{\beta,\text{emd}}h\right)^{1/2}. \tag{8.36}$$

8.11 Hausman Equality

Form (8.25) we have

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right) &= \widehat{\boldsymbol{V}}_\beta \boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta \boldsymbol{R}\right)^{-1}\sqrt{n}\left(\boldsymbol{R}'\widehat{\beta}_{\text{ols}} - \boldsymbol{c}\right) \\ &\underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_\beta \boldsymbol{R}\right)^{-1}\boldsymbol{R}'\boldsymbol{V}_\beta\right).\end{aligned}$$

It follows that the asymptotic variances of the estimators satisfy the relationship

$$\text{avar}\left[\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right] = \text{avar}\left[\widehat{\beta}_{\text{ols}}\right] - \text{avar}\left[\widetilde{\beta}_{\text{emd}}\right]. \tag{8.37}$$

We call (8.37) the **Hausman Equality**: the asymptotic variance of the difference between an efficient and another estimator is the difference in the asymptotic variances.

8.12 Example: Mankiw, Romer and Weil (1992)

We illustrate the methods by replicating some of the estimates reported in a well-known paper by Mankiw, Romer, and Weil (1992). The paper investigates the implications of the Solow growth model using cross-country regressions. A key equation in their paper regresses the change between 1960 and 1985 in log GDP per capita on (1) log GDP in 1960, (2) the log of the ratio of aggregate investment to

Table 8.1: Estimates of Solow Growth Model

	$\widehat{\beta}_{\text{ols}}$	$\widehat{\beta}_{\text{cls}}$	$\widehat{\beta}_{\text{emd}}$
$\log GDP_{1960}$	−0.29	−0.30	−0.30
	(0.05)	(0.05)	(0.05)
$\log \frac{I}{GDP}$	0.52	0.50	0.46
	(0.11)	(0.09)	(0.08)
$\log\left(n+g+\delta\right)$	−0.51	−0.74	−0.71
	(0.24)	(0.08)	(0.07)
log(School)	0.23	0.24	0.25
	(0.07)	(0.07)	(0.06)
Intercept	3.02	2.46	2.48
	(0.74)	(0.44)	(0.44)

Standard errors are heteroskedasticity-consistent

GDP, (3) the log of the sum of the population growth rate n, the technological growth rate g, and the rate of depreciation δ, and (4) the log of the percentage of the working-age population that is in secondary school (*School*), the latter a proxy for human-capital accumulation.

The data is available on the textbook webpage in the file `MRW1992`.

The sample is 98 non-oil-producing countries and the data was reported in the published paper. As g and δ were unknown the authors set $g+\delta = 0.05$. We report least squares estimates in the first column of Table 8.1. The estimates are consistent with the Solow theory due to the positive coefficients on investment and human capital and negative coefficient for population growth. The estimates are also consistent with the convergence hypothesis (that income levels tend towards a common mean over time) as the coefficient on intial GDP is negative.

The authors show that in the Solow model the 2^{nd}, 3^{rd} and 4^{th} coefficients sum to zero. They reestimated the equation imposing this constraint. We present constrained least squares estimates in the second column of Table 8.1 and efficient minimum distance estimates in the third column. Most of the coefficients and standard errors only exhibit small changes by imposing the constraint. The one exception is the coefficient on log population growth which increases in magnitude and its standard error decreases substantially. The differences between the CLS and EMD estimates are modest.

We now present Stata, R and MATLAB code which implements these estimates.

You may notice that the Stata code has a section which uses the Mata matrix programming language. This is used because Stata does not implement the efficient minimum distance estimator, so needs to be separately programmed. As illustrated here, the Mata language allows a Stata user to implement methods using commands which are quite similar to MATLAB.

Stata do File

```
use "MRW1992.dta", clear
gen lndY = log(Y85)-log(Y60)
gen lnY60 = log(Y60)
gen lnI = log(invest/100)
gen lnG = log(pop_growth/100+0.05)
gen lnS = log(school/100)
* Unrestricted regression
reg lndY lnY60 lnI lnG lnS if N==1, r
* Store result for efficient minimum distance
mat b = e(b)'
scalar k = e(rank)
mat V = e(V)
* Constrained regression
constraint define 1 lnI+lnG+lnS=0
cnsreg lndY lnY60 lnI lnG lnS if N==1, constraints(1) r
* Efficient minimum distance
mata{
      data = st_data(.,("lnY60","lnI","lnG","lnS","lndY","N"))
      data_select = select(data,data[.,6]:==1)
      y = data_select[.,5]
      n = rows(y)
      x = (data_select[.,1..4],J(n,1,1))
      k = cols(x)
      invx = invsym(x'*x)
      b_ols = st_matrix("b")
      V_ols = st_matrix("V")
      R = (0 \ 1 \ 1 \ 1 \ 0)
      b_emd = b_ols-V_ols*R*invsym(R'*V_ols*R)*R'*b_ols
      e_emd = J(1,k,y-x*b_emd)
      xe_emd = x:*e_emd
      xe_emd'*xe_emd
      V2 = (n/(n-k+1))*invx*(xe_emd'*xe_emd)*invx
      V_emd = V2 - V2*R*invsym(R'*V2*R)*R'*V2
      se_emd = diagonal(sqrt(V_emd))
      st_matrix("b_emd",b_emd)
      st_matrix("se_emd",se_emd)}
mat list b_emd
mat list se_emd
```

R Program File

```
data <- read.table("MRW1992.txt",header=TRUE)
N <- matrix(data$N,ncol=1)
lndY <- matrix(log(data$Y85)-log(data$Y60),ncol=1)
lnY60 <- matrix(log(data$Y60),ncol=1)
lnI <- matrix(log(data$invest/100),ncol=1)
lnG <- matrix(log(data$pop_growth/100+0.05),ncol=1)
lnS <- matrix(log(data$school/100),ncol=1)
xx <- as.matrix(cbind(lnY60,lnI,lnG,lnS,matrix(1,nrow(lndY),1)))
x <- xx[N==1,]
y <- lndY[N==1]
n <- nrow(x)
k <- ncol(x)
# Unrestricted regression
invx <-solve(t(x)%*%x)
b_ols <- solve((t(x)%*%x),(t(x)%*%y))
e_ols <- rep((y-x%*%beta_ols),times=k)
xe_ols <- x*e_ols
V_ols <- (n/(n-k))*invx%*%(t(xe_ols)%*%xe_ols)%*%invx
se_ols <- sqrt(diag(V_ols))
print(beta_ols)
print(se_ols)
# Constrained regression
R <- c(0,1,1,1,0)
iR <- invx%*%R%*%solve(t(R)%*%invx%*%R)%*%t(R)
b_cls <- b_ols - iR%*%b_ols
e_cls <- rep((y-x%*%b_cls),times=k)
xe_cls <- x*e_cls
V_tilde <- (n/(n-k+1))*invx%*%(t(xe_cls)%*%xe_cls)%*%invx
V_cls <- V_tilde - iR%*%V_tilde - V_tilde%*%t(iR) +iR%*%V_tilde%*%t(iR)
print(b_cls)print(se_cls)
# Efficient minimum distance
Vr <- V_ols%*%R%*%solve(t(R)%*%V_ols%*%R)%*%t(R)
b_emd <- b_ols - Vr%*%b_ols
e_emd <- rep((y-x%*%b_emd),times=k)
xe_emd <- x*e_emd
V2 <- (n/(n-k+1))*invx%*%(t(xe_emd)%*%xe_emd)%*%invx
V_emd <- V2 - V2%*%R%*%solve(t(R)%*%V2%*%R)%*%t(R)%*%V2
se_emd <- sqrt(diag(V_emd))
```

MATLAB Program File

```
data = xlsread('MRW1992.xlsx');
N = data(:,1);
Y60 = data(:,4);
Y85 = data(:,5);
pop_growth = data(:,7);
invest = data(:,8);
school = data(:,9);
lndY = log(Y85)-log(Y60);
lnY60 = log(Y60);
lnI = log(invest/100);
lnG = log(pop_growth/100+0.05);
lnS = log(school/100);
xx = [lnY60,lnI,lnG,lnS,ones(size(lndY,1),1)];
x = xx(N==1,:);
y = lndY(N==1);
[n,k] = size(x);
% Unrestricted regression
invx = inv(x'*x);
beta_ols = (x'*x)\(x'*y);
xe_ols = x.*(y-x*beta_ols);
V_ols = (n/(n-k))*invx*(xe_ols'*xe_ols)*invx;
se_ols = sqrt(diag(V_ols));
display(beta_ols);
display(se_ols);
% Constrained regression
R = [0;1;1;1;0];
iR = invx*R*inv(R'*invx*R)*R';
beta_cls = beta_ols - iR*beta_ols;
xe_cls = x.*(y-x*beta_cls);
V_tilde = (n/(n-k+1))*invx*(xe_cls'*xe_cls)*invx;
V_cls = V_tilde - iR*V_tilde - V_tilde*(iR') + iR*V_tilde*(iR');
se_cls = sqrt(diag(V_cls));
display(beta_cls);display(se_cls);
% Efficient minimum distance
beta_emd = beta_ols-V_ols*R*inv(R'*V_ols*R)*R'*beta_ols;
xe_emd = x.*(y-x*beta_emd);
V2 = (n/(n-k+1))*invx*(xe_emd'*xe_emd)*invx;
V_emd = V2 - V2*R*inv(R'*V2*R)*R'*V2;
se_emd = sqrt(diag(V_emd));
display(beta_emd);display(se_emd);
```

8.13 Misspecification

What are the consequences for a constrained estimator $\widetilde{\beta}$ if the constraint (8.1) is incorrect? To be specific suppose that the truth is

$$\boldsymbol{R}'\beta = \boldsymbol{c}^*$$

where $\boldsymbol{c}^*$ is not necessarily equal to $\boldsymbol{c}$.

This situation is a generalization of the analysis of "omitted variable bias" from Section 2.24 where we found that the short regression (e.g. (8.12)) is estimating a different projection coefficient than the long regression (e.g. (8.10)).

One answer is to apply formula (8.23) to find that

$$\widetilde{\beta}_{\text{md}} \xrightarrow[p]{} \beta^*_{\text{md}} = \beta - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{c}^* - \boldsymbol{c}\right). \tag{8.38}$$

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 β^*_{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

$$\beta^*_{\text{cls}} = \underset{\boldsymbol{R}'\beta=\boldsymbol{c}}{\operatorname{argmin}} \mathbb{E}\left[\left(Y - X'\beta\right)^2\right],$$

the best linear predictor subject to the constraint (8.1). The minimum distance estimator converges in probability to

$$\beta^*_{\text{md}} = \underset{\boldsymbol{R}'\beta=\boldsymbol{c}}{\operatorname{argmin}} \left(\beta - \beta_0\right)' \boldsymbol{W}\left(\beta - \beta_0\right)$$

where β_0 is the true coefficient. That is, β^*_{md} is the coefficient vector satisfying (8.1) closest to the true value in the weighted Euclidean norm. These calculations show that the constrained estimators are still reasonable in the sense that they produce good approximations to the true coefficient conditional on being required to satisfy the constraint.

We can also show that $\widetilde{\beta}_{\text{md}}$ has an asymptotic normal distribution. The trick is to define the pseudo-true value

$$\beta^*_n = \beta - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{c}^* - \boldsymbol{c}\right). \tag{8.39}$$

(Note that (8.38) and (8.39) are different!) Then

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\beta}_{\text{md}} - \beta^*_n\right) &= \sqrt{n}\left(\widehat{\beta} - \beta\right) - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\sqrt{n}\left(\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c}^*\right) \\
&= \left(\boldsymbol{I}_k - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\beta} - \beta\right) \\
&\xrightarrow[d]{} \left(\boldsymbol{I}_k - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\mathrm{N}\left(0, \boldsymbol{V}_\beta\right) \\
&= \mathrm{N}\left(0, \boldsymbol{V}_\beta(\boldsymbol{W})\right).
\end{aligned} \tag{8.40}$$

In particular

$$\sqrt{n}\left(\widetilde{\beta}_{\text{emd}} - \beta^*_n\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}^*_\beta\right).$$

This means that even when the constraint (8.1) is misspecified the conventional covariance matrix estimator (8.35) and standard errors (8.36) are appropriate measures of the sampling variance though the

distributions are centered at the pseudo-true values (projections) β_n^* rather than β. 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 β_n by n via the relationship

$$\boldsymbol{R}'\beta_n = \boldsymbol{c} + \delta n^{-1/2}. \tag{8.41}$$

for some $\delta \in \mathbb{R}^q$. Equation (8.41) specifies that β_n violates (8.1) and thus the constraint is misspecified. However, the constraint is "close" to correct as the difference $\boldsymbol{R}'\beta_n - \boldsymbol{c} = \delta n^{-1/2}$ is "small" in the sense that it decreases with the sample size n. We call (8.41) **local misspecification**.

The asymptotic theory is derived as $n \to \infty$ under the sequence of probability distributions with the coefficients β_n. The way to think about this is that the true value of the parameter is β_n and it is "close" to satisfying (8.1). The reason why the deviation is proportional to $n^{-1/2}$ is because this is the only choice under which the localizing parameter δ appears in the asymptotic distribution but does not dominate it. The best way to see this is to work through the asymptotic approximation.

Since β_n is the true coefficient value, then $Y = X'\beta_n + e$ and we have the standard representation for the unconstrained estimator, namely

$$\sqrt{n}\left(\widehat{\beta} - \beta_n\right) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i e_i\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right). \tag{8.42}$$

There is no difference under fixed (classical) or local asymptotics since the right-hand-side is independent of the coefficient β_n.

A difference arises for the constrained estimator. Using (8.41), $\boldsymbol{c} = \boldsymbol{R}'\beta_n - \delta n^{-1/2}$ so

$$\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c} = \boldsymbol{R}'\left(\widehat{\beta} - \beta_n\right) + \delta n^{-1/2}$$

and

$$\begin{aligned}\widetilde{\beta}_{\mathrm{md}} &= \widehat{\beta} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c}\right) \\ &= \widehat{\beta} - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\left(\widehat{\beta} - \beta_n\right) + \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\delta n^{-1/2}.\end{aligned}$$

It follows that

$$\sqrt{n}\left(\widetilde{\beta}_{\mathrm{md}} - \beta_n\right) = \left(\boldsymbol{I}_k - \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\sqrt{n}\left(\widehat{\beta} - \beta_n\right) + \widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{W}}^{-1}\boldsymbol{R}\right)^{-1}\delta.$$

The first term is asymptotically normal (from 8.42)). The second term converges in probability to a constant. This is because the $n^{-1/2}$ local scaling in (8.41) 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

$$\sqrt{n}\left(\widetilde{\beta}_{\mathrm{md}} - \beta_n\right) \underset{d}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{V}_\beta\right) + \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\delta = \mathrm{N}\left(\delta^*, \boldsymbol{V}_\beta(\boldsymbol{W})\right) \tag{8.43}$$

where $\delta^* = \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\delta$.

The asymptotic distribution (8.43) is an approximation of the sampling distribution of the restricted estimator under misspecification. The distribution (8.43) contains an asymptotic bias component δ^*. The approximation is not fundamentally different from (8.40) – they both have the same asymptotic variances and both reflect the bias due to misspecification. The difference is that (8.40) puts the bias on the left-side of the convergence arrow while (8.43) has the bias on the right-side. There is no substantive difference between the two. However, (8.43) is more convenient for some purposes such as the analysis of the power of tests as we will explore in the next chapter.

8.14 Nonlinear Constraints

In some cases it is desirable to impose nonlinear constraints on the parameter vector β. They can be written as

$$r(\beta) = 0 \tag{8.44}$$

where $r : \mathbb{R}^k \to \mathbb{R}^q$. This includes the linear constraints (8.1) as a special case. An example of (8.44) which cannot be written as (8.1) is $\beta_1\beta_2 = 1$, which is (8.44) with $r(\beta) = \beta_1\beta_2 - 1$.

The constrained least squares and minimum distance estimators of β subject to (8.44) solve the minimization problems

$$\widetilde{\beta}_{\text{cls}} = \underset{r(\beta)=0}{\operatorname{argmin}} \operatorname{SSE}(\beta) \tag{8.45}$$

$$\widetilde{\beta}_{\text{md}} = \underset{r(\beta)=0}{\operatorname{argmin}}\ J\left(\beta\right) \tag{8.46}$$

where $\operatorname{SSE}(\beta)$ and $J\left(\beta\right)$ are defined in (8.4) and (8.19), respectively. The solutions solve the Lagrangians

$$\mathscr{L}(\beta, \lambda) = \frac{1}{2}\operatorname{SSE}(\beta) + \lambda' r(\beta)$$

or

$$\mathscr{L}(\beta, \lambda) = \frac{1}{2} J\left(\beta\right) + \lambda' r(\beta) \tag{8.47}$$

over (β, λ).

Computationally there is no general closed-form solution so they must be found numerically. Algorithms to numerically solve (8.45) and (8.46) are known as **constrained optimization** methods and are available in programming languages including MATLAB and R. See Chapter 12 of *Probability and Statistics for Economists*.

Assumption 8.3

1. $r(\beta) = 0$.
2. $r(\beta)$ is continuously differentiable at the true β.
3. $\operatorname{rank}(\boldsymbol{R}) = q$, where $\boldsymbol{R} = \dfrac{\partial}{\partial \beta} r(\beta)'$.

The asymptotic distribution is a simple generalization of the case of a linear constraint but the proof is more delicate.

Theorem 8.10 Under Assumptions 7.2, 8.2, and 8.3, for $\widetilde{\beta} = \widetilde{\beta}_{\text{md}}$ and $\widetilde{\beta} = \widetilde{\beta}_{\text{cls}}$ defined in (8.45) and (8.46),

$$\sqrt{n}\left(\widetilde{\beta} - \beta\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_\beta(\boldsymbol{W})\right)$$

as $n \to \infty$ where $\boldsymbol{V}_\beta(\boldsymbol{W})$ is defined in (8.24). For $\widetilde{\beta}_{\text{cls}}$, $\boldsymbol{W} = \boldsymbol{Q}_{XX}$ and $\boldsymbol{V}_\beta(\boldsymbol{W}) = \boldsymbol{V}_{\text{cls}}$ as defined in Theorem 8.8. $\boldsymbol{V}_\beta(\boldsymbol{W})$ is minimized with $\boldsymbol{W} = \boldsymbol{V}_\beta^{-1}$ in which case the asymptotic variance is

$$\boldsymbol{V}_\beta^* = \boldsymbol{V}_\beta - \boldsymbol{V}_\beta \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_\beta \boldsymbol{R}\right)^{-1} \boldsymbol{R}'\boldsymbol{V}_\beta.$$

The asymptotic covariance matrix for the efficient minimum distance estimator can be estimated by

$$\widehat{\boldsymbol{V}}_\beta^* = \widehat{\boldsymbol{V}}_\beta - \widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta$$

where

$$\widehat{\boldsymbol{R}} = \frac{\partial}{\partial \beta} r(\widetilde{\beta}_{\text{md}})'. \tag{8.48}$$

Standard errors for the elements of $\widetilde{\beta}_{\text{md}}$ are the square roots of the diagonal elements of $\widehat{\boldsymbol{V}}_{\widetilde{\beta}}^* = n^{-1}\widehat{\boldsymbol{V}}_\beta^*$.

8.15 Inequality Restrictions

Inequality constraints on the parameter vector β take the form

$$r(\beta) \geq 0 \tag{8.49}$$

for some function $r : \mathbb{R}^k \to \mathbb{R}^q$. 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

$$\widetilde{\beta}_{\text{cls}} = \underset{r(\beta)\geq 0}{\operatorname{argmin}} \operatorname{SSE}(\beta) \tag{8.50}$$

and

$$\widetilde{\beta}_{\text{md}} = \underset{r(\beta)\geq 0}{\operatorname{argmin}}\ J\left(\beta\right). \tag{8.51}$$

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 the following approach. Compute the uncontrained estimator $\widehat{\beta}$. If $\widehat{\beta}_1 \geq 0$ then $\widetilde{\beta} = \widehat{\beta}$. Otherwise if $\widehat{\beta}_1 < 0$ then impose $\beta_1 = 0$ (eliminate the regressor X_1) and re-estimate. This method 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 computation problems (8.50) and (8.51) are examples of **quadratic programming**. Quick computer algorithms are available in programming languages including MATLAB 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 $r(\beta) > 0$ 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., $r(\beta) = 0$, then the estimator has a truncated structure. This is easiest to see in the one-dimensional case. If we have an estimator $\widehat{\beta}$ which satisfies $\sqrt{n}\left(\widehat{\beta} - \beta\right) \underset{d}{\longrightarrow} Z = \mathrm{N}\left(0, V_\beta\right)$ and $\beta = 0$, then the constrained estimator $\widetilde{\beta} = \max[\widehat{\beta}, 0]$ will have the asymptotic distribution $\sqrt{n}\widetilde{\beta} \underset{d}{\longrightarrow} \max[Z, 0]$, a "half-normal" distribution.

8.16 Technical Proofs*

Proof of Theorem 8.9, equation (8.28) Let $\boldsymbol{R}_\perp$ be a full rank $k \times \left(k - q\right)$ matrix satisfying $\boldsymbol{R}_\perp' \boldsymbol{V}_\beta \boldsymbol{R} = 0$ and then set $\boldsymbol{C} = [\boldsymbol{R}, \boldsymbol{R}_\perp]$ which is full rank and invertible. Then we can calculate that

$$\boldsymbol{C}' \boldsymbol{V}_\beta^* \boldsymbol{C} = \begin{bmatrix} \boldsymbol{R}' \boldsymbol{V}_\beta^* \boldsymbol{R} & \boldsymbol{R}' \boldsymbol{V}_\beta^* \boldsymbol{R}_\perp \\ \boldsymbol{R}_\perp' \boldsymbol{V}_\beta^* \boldsymbol{R} & \boldsymbol{R}_\perp' \boldsymbol{V}_\beta^* \boldsymbol{R}_\perp \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{R}_\perp' \boldsymbol{V}_\beta \boldsymbol{R}_\perp \end{bmatrix}$$

and

$$\begin{aligned} &\boldsymbol{C}' \boldsymbol{V}_\beta(\boldsymbol{W}) \boldsymbol{C} \\ &= \begin{bmatrix} \boldsymbol{R}' \boldsymbol{V}_\beta^*(\boldsymbol{W}) \boldsymbol{R} & \boldsymbol{R}' \boldsymbol{V}_\beta^*(\boldsymbol{W}) \boldsymbol{R}_\perp \\ \boldsymbol{R}_\perp' \boldsymbol{V}_\beta^*(\boldsymbol{W}) \boldsymbol{R} & \boldsymbol{R}_\perp' \boldsymbol{V}_\beta^*(\boldsymbol{W}) \boldsymbol{R}_\perp \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{R}_\perp' \boldsymbol{V}_\beta \boldsymbol{R}_\perp + \boldsymbol{R}_\perp' \boldsymbol{W} \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}\right)^{-1} \boldsymbol{R}' \boldsymbol{V}_\beta \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}\right)^{-1} \boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}_\perp \end{bmatrix}. \end{aligned}$$

Thus

$$\begin{aligned} &\boldsymbol{C}' \left(\boldsymbol{V}_\beta(\boldsymbol{W}) - \boldsymbol{V}_\beta^*\right) \boldsymbol{C} \\ &= \boldsymbol{C}' \boldsymbol{V}_\beta(\boldsymbol{W}) \boldsymbol{C} - \boldsymbol{C}' \boldsymbol{V}_\beta^* \boldsymbol{C} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & \boldsymbol{R}_\perp' \boldsymbol{W} \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}\right)^{-1} \boldsymbol{R}' \boldsymbol{V}_\beta \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}\right)^{-1} \boldsymbol{R}' \boldsymbol{W} \boldsymbol{R}_\perp \end{bmatrix} \\ &\geq 0 \end{aligned}$$

Since $\boldsymbol{C}$ is invertible it follows that $\boldsymbol{V}_\beta(\boldsymbol{W}) - \boldsymbol{V}_\beta^* \geq 0$ which is (8.28). ■

Proof of Theorem 8.10 We show the result for the minimum distance estimator $\widetilde{\beta} = \widetilde{\beta}_{\mathrm{md}}$ as the proof for the constrained least squares estimator is similar. For simplicity we assume that the constrained estimator is consistent $\widetilde{\beta} \underset{p}{\longrightarrow} \beta$. This can be shown with more effort, but requires a deeper treatment than appropriate for this textbook.

For each element $r_j(\beta)$ of the q-vector $r(\beta)$, by the mean value theorem there exists a β_j^* on the line segment joining $\widetilde{\beta}$ and β such that

$$r_j(\widetilde{\beta}) = r_j(\beta) + \frac{\partial}{\partial \beta} r_j(\beta_j^*)' \left(\widetilde{\beta} - \beta\right). \tag{8.52}$$

Let $\boldsymbol{R}_n^*$ be the $k \times q$ matrix

$$\boldsymbol{R}^* = \begin{bmatrix} \dfrac{\partial}{\partial \beta} r_1(\beta_1^*) & \dfrac{\partial}{\partial \beta} r_2(\beta_2^*) & \cdots & \dfrac{\partial}{\partial \beta} r_q(\beta_q^*) \end{bmatrix}.$$

Since $\widetilde{\beta} \underset{p}{\longrightarrow} \beta$ it follows that $\beta_j^* \underset{p}{\longrightarrow} \beta$, and by the CMT, $\boldsymbol{R}^* \underset{p}{\longrightarrow} \boldsymbol{R}$. Stacking the (8.52), we obtain

$$r(\widetilde{\beta}) = r(\beta) + \boldsymbol{R}^{*\prime}\left(\widetilde{\beta} - \beta\right).$$

Since $r\left(\widetilde{\beta}\right) = 0$ by construction and $r(\beta) = 0$ by Assumption 8.1 this implies

$$0 = \boldsymbol{R}^{*\prime}\left(\widetilde{\beta} - \beta\right). \tag{8.53}$$

The first-order condition for (8.47) is $\widehat{\boldsymbol{W}}\left(\widehat{\beta} - \widetilde{\beta}\right) = \widehat{\boldsymbol{R}}\widetilde{\lambda}$ where $\widehat{\boldsymbol{R}}$ is defined in (8.48). Premultiplying by $\boldsymbol{R}^{*\prime}\widehat{\boldsymbol{W}}^{-1}$, inverting, and using (8.53), we find

$$\widetilde{\lambda} = \left(\boldsymbol{R}^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}^{*\prime}\left(\widehat{\beta} - \widetilde{\beta}\right) = \left(\boldsymbol{R}^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}^{*\prime}\left(\widehat{\beta} - \beta\right).$$

Thus

$$\widetilde{\beta} - \beta = \left(\boldsymbol{I}_k - \widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\left(\boldsymbol{R}_n^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}_n^{*\prime}\right)\left(\widehat{\beta} - \beta\right). \tag{8.54}$$

From Theorem 7.3 and Theorem 7.6 we find

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\beta} - \beta\right) &= \left(\boldsymbol{I}_k - \widehat{\boldsymbol{W}}^{-1}\widehat{\boldsymbol{R}}\left(\boldsymbol{R}_n^{*\prime}\widehat{\boldsymbol{W}}^{-1}\widetilde{\boldsymbol{R}}\right)^{-1}\boldsymbol{R}_n^{*\prime}\right)\sqrt{n}\left(\widehat{\beta} - \beta\right) \\
&\underset{d}{\longrightarrow} \left(\boldsymbol{I}_k - \boldsymbol{W}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{W}^{-1}\boldsymbol{R}\right)^{-1}\boldsymbol{R}'\right)\mathrm{N}\left(0, \boldsymbol{V}_\beta\right) \\
&= \mathrm{N}\left(0, \boldsymbol{V}_\beta(\boldsymbol{W})\right).
\end{aligned}$$

■

8.17 Exercises

Exercise 8.1 In the model $Y = X_1'\beta_1 + X_2'\beta_2 + e$, show directly from definition (8.3) that the CLS estimator of $\beta = (\beta_1, \beta_2)$ subject to the constraint that $\beta_2 = 0$ is the OLS regression of Y on X_1.

Exercise 8.2 In the model $Y = X_1'\beta_1 + X_2'\beta_2 + e$, show directly from definition (8.3) that the CLS estimator of $\beta = (\beta_1, \beta_2)$ subject to the constraint $\beta_1 = \boldsymbol{c}$ (where $\boldsymbol{c}$ is some given vector) is OLS of $Y - X_1'\boldsymbol{c}$ on X_2.

Exercise 8.3 In the model $Y = X_1'\beta_1 + X_2'\beta_2 + e$, with β_1 and β_2 each $k \times 1$, find the CLS estimator of $\beta = (\beta_1, \beta_2)$ subject to the constraint that $\beta_1 = -\beta_2$.

Exercise 8.4 In the linear projection model $Y = \alpha + X'\beta + e$ consider the restriction $\beta = 0$.

(a) Find the CLS estimator of α under the restriction $\beta = 0$.

(b) Find an expression for the efficient minimum distance estimator of α under the restriction $\beta = 0$.

Exercise 8.5 Verify that for $\widetilde{\beta}_{\text{cls}}$ defined in (8.8) that $\boldsymbol{R}'\widetilde{\beta}_{\text{cls}} = \boldsymbol{c}$.

Exercise 8.6 Prove Theorem 8.1.

Exercise 8.7 Prove Theorem 8.2, that is, $\mathbb{E}\left[\widetilde{\beta}_{\text{cls}} \mid \boldsymbol{X}\right] = \beta$, under the assumptions of the linear regression regression model and (8.1). (Hint: Use Theorem 8.1.)

Exercise 8.8 Prove Theorem 8.3.

Exercise 8.9 Prove Theorem 8.4. That is, show $\mathbb{E}\left[s^2_{\text{cls}} \mid \boldsymbol{X}\right] = \sigma^2$ under the assumptions of the homoskedastic regression model and (8.1).

Exercise 8.10 Verify (8.22), (8.23), and that the minimum distance estimator $\widetilde{\beta}_{\text{md}}$ with $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{XX}$ equals the CLS estimator.

Exercise 8.11 Prove Theorem 8.6.

Exercise 8.12 Prove Theorem 8.7.

Exercise 8.13 Prove Theorem 8.8. (Hint: Use that CLS is a special case of Theorem 8.7.)

Exercise 8.14 Verify that (8.26) is $\boldsymbol{V}_\beta(\boldsymbol{W})$ with $\boldsymbol{W} = \boldsymbol{V}_\beta^{-1}$.

Exercise 8.15 Prove (8.27). Hint: Use (8.26).

Exercise 8.16 Verify (8.29), (8.30) and (8.31).

Exercise 8.17 Verify (8.32), (8.33), and (8.34).

Exercise 8.18 Suppose you have two independent samples each with n observations which satisfy the models $Y_1 = X_1'\beta_1 + e_1$ with $\mathbb{E}[X_1 e_1] = 0$ and $Y_2 = X_2'\beta_2 + e_2$ with $\mathbb{E}[X_2 e_2] = 0$ where β_1 and β_2 are both $k \times 1$. You estimate β_1 and β_2 by OLS on each sample, with consistent asymptotic covariance matrix estimators $\widehat{\boldsymbol{V}}_{\beta_1}$ and $\widehat{\boldsymbol{V}}_{\beta_2}$. Consider efficient minimum distance estimation under the restriction $\beta_1 = \beta_2$.

(a) Find the estimator $\widetilde{\beta}$ of $\beta = \beta_1 = \beta_2$.

(b) Find the asymptotic distribution of $\widetilde{\beta}$.

(c) How would you approach the problem if the sample sizes are different, say n_1 and n_2?

Exercise 8.19 Use the `cps09mar` dataset and the subsample of white male Hispanics.

(a) Estimate the regression

$$\begin{aligned}\widehat{\log(\mathit{wage})} = {} & \beta_1\, \mathit{education} + \beta_2\, \mathit{experience} + \beta_3\, \mathit{experience}^2/100 + \beta_4 \mathit{married}_1 \\ & + \beta_5 \mathit{married}_2 + \beta_6 \mathit{married}_3 + \beta_7 \mathit{widowed} + \beta_8 \mathit{divorced} + \beta_9 \mathit{separated} + \beta_{10}\end{aligned}$$

where $\mathit{married}_1$, $\mathit{married}_2$, and $\mathit{married}_3$ are the first three marital codes listed in Section 3.22.

(b) Estimate the equation by CLS imposing the constraints $\beta_4 = \beta_7$ and $\beta_8 = \beta_9$. Report the estimates and standard errors.

(c) Estimate the equation using efficient minimum distance imposing the same constraints. Report the estimates and standard errors.

(d) Under what constraint on the coefficients is the wage equation non-decreasing in experience for experience up to 50?

(e) Estimate the equation imposing $\beta_4 = \beta_7$, $\beta_8 = \beta_9$, and the inequality from part (d).

Exercise 8.20 Take the model

$$\begin{aligned} Y &= m(X) + e \\ m(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p \\ \mathbb{E}\left[X^j e\right] &= 0, \qquad j = 0, ..., p \\ g(x) &= \frac{d}{dx} m(x) \end{aligned}$$

with i.i.d. observations (Y_i, X_i), $i = 1, ..., n$. The order of the polynomial p is known.

(a) How should we interpret the function $m(x)$ given the projection assumption? How should we interpret $g(x)$? (Briefly)

(b) Describe an estimator $\widehat{g}(x)$ of $g(x)$.

(c) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{g}(x) - g(x)\right)$ as $n \to \infty$.

(d) Show how to construct an asymptotic 95% confidence interval for $g(x)$ (for a single x).

(e) Assume $p = 2$. Describe how to estimate $g(x)$ imposing the constraint that $m(x)$ is concave.

(f) Assume $p = 2$. Describe how to estimate $g(x)$ imposing the constraint that $m(u)$ is increasing on the region $u \in [x_L, x_U]$.

Exercise 8.21 Take the linear model with restrictions $Y = X'\beta + e$ with $\mathbb{E}[Xe] = 0$ and $\boldsymbol{R}'\beta = \boldsymbol{c}$. Consider three estimators for β:

- $\widehat{\beta}$ the unconstrained least squares estimator
- $\widetilde{\beta}$ the constrained least squares estimator
- $\overline{\beta}$ the constrained efficient minimum distance estimator

For the three estimator define the residuals $\widehat{e}_i = Y_i - X_i'\widehat{\beta}$, $\widetilde{e}_i = Y_i - X_i'\widetilde{\beta}$, $\overline{e}_i = Y_i - X_i'\overline{\beta}$, and variance estimators $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \widehat{e}_i^2$, $\widetilde{\sigma}^2 = n^{-1}\sum_{i=1}^n \widetilde{e}_i^2$, and $\overline{\sigma}^2 = n^{-1}\sum_{i=1}^n \overline{e}_i^2$.

(a) As $\overline{\beta}$ is the most efficient estimator and $\widehat{\beta}$ the least, do you expect $\overline{\sigma}^2 < \widetilde{\sigma}^2 < \widehat{\sigma}^2$ in large samples?

(b) Consider the statistic

$$T_n = \widehat{\sigma}^{-2} \sum_{i=1}^n (\widehat{e}_i - \widetilde{e}_i)^2.$$

Find the asymptotic distribution for T_n when $\boldsymbol{R}'\beta = \boldsymbol{c}$ is true.

(c) Does the result of the previous question simplify when the error e_i is homoskedastic?

Exercise 8.22 Take the linear model $Y = X_1\beta_1 + X_2\beta_2 + e$ with $\mathbb{E}[Xe] = 0$. Consider the restriction $\dfrac{\beta_1}{\beta_2} = 2$.

(a) Find an explicit expression for the CLS estimator $\widetilde{\beta} = (\widetilde{\beta}_1, \widetilde{\beta}_2)$ of $\beta = (\beta_1, \beta_2)$ under the restriction. Your answer should be specific to the restriction. It should not be a generic formula for an abstract general restriction.

(b) Derive the asymptotic distribution of $\widetilde{\beta}_1$ under the assumption that the restriction is true.

Chapter 9

Hypothesis Testing

In Chapter 5 we briefly introduced hypothesis testing in the context of the normal regression model. In this chapter we explore hypothesis testing in greater detail with a particular emphasis on asymptotic inference. For more detail on the foundations see Chapter 13 of *Probability and Statistics for Economists*.

9.1 Hypotheses

In Chapter 8 we discussed estimation subject to restrictions, including linear restrictions (8.1), nonlinear restrictions (8.44), and inequality restrictions (8.49). In this chapter we discuss **tests** of such restrictions.

Hypothesis tests attempt to assess whether there is evidence contrary to a proposed restriction. Let $\theta = r(\beta)$ be a $q \times 1$ parameter of interest where $r : \mathbb{R}^k \to \Theta \subset \mathbb{R}^q$ is some transformation. For example, θ may be a single coefficient, e.g. $\theta = \beta_j$, the difference between two coefficients, e.g. $\theta = \beta_j - \beta_\ell$, or the ratio of two coefficients, e.g. $\theta = \beta_j / \beta_\ell$.

A point hypothesis concerning θ is a proposed restriction such as

$$\theta = \theta_0 \tag{9.1}$$

where θ_0 is a hypothesized (known) value.

More generally, letting $\beta \in B \subset \mathbb{R}^k$ be the parameter space, a hypothesis is a restriction $\beta \in B_0$ where B_0 is a proper subset of B. This specializes to (9.1) by setting $B_0 = \{\beta \in B : r(\beta) = \theta_0\}$.

In this chapter we will focus exclusively on point hypotheses of the form (9.1) as they are the most common and relatively simple to handle.

The hypothesis to be tested is called the null hypothesis.

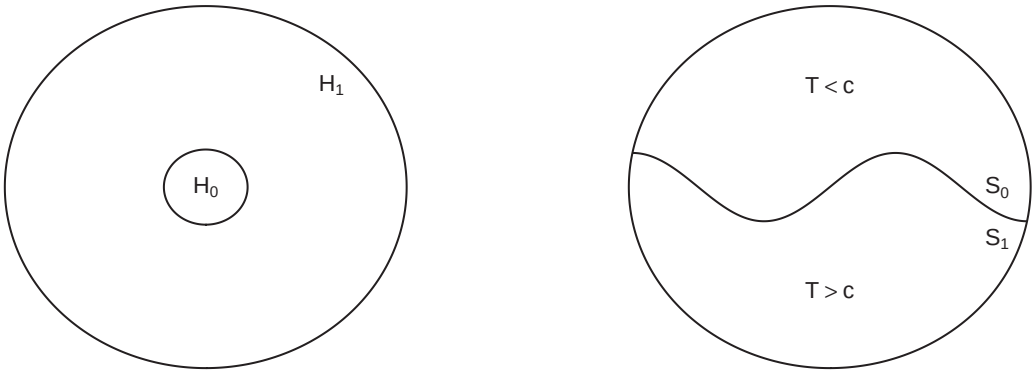
Definition 9.1 The **null hypothesis** $\mathbb{H}_0$ is the restriction $\theta = \theta_0$ or $\beta \in B_0$.

We often write the null hypothesis as $\mathbb{H}_0 : \theta = \theta_0$ or $\mathbb{H}_0 : r(\beta) = \theta_0$.

The complement of the null hypothesis (the collection of parameter values which do not satisfy the null hypothesis) is called the alternative hypothesis.

Definition 9.2 The **alternative hypothesis** $\mathbb{H}_1$ is the set $\{\theta \in \Theta : \theta \neq \theta_0\}$ or $\{\beta \in B : \beta \notin B_0\}$.

We often write the alternative hypothesis as $\mathbb{H}_1 : \theta \neq \theta_0$ or $\mathbb{H}_1 : r(\beta) \neq \theta_0$. For simplicity, we often refer to the hypotheses as "the null" and "the alternative". Figure 9.1(a) illustrates the division of the parameter space into null and alternative hypotheses.



(a) Null and Alternative Hypotheses (b) Acceptance and Rejection Regions

Figure 9.1: Hypothesis Testing

In hypothesis testing, we assume that there is a true (but unknown) value of θ and this value either satisfies $\mathbb{H}_0$ or does not satisfy $\mathbb{H}_0$. The goal of hypothesis testing is to assess whether or not $\mathbb{H}_0$ is true by asking if $\mathbb{H}_0$ is consistent with the observed data.

To be specific, take our example of wage determination and consider the question: Does union membership affect wages? We can turn this into a hypothesis test by specifying the null as the restriction that a coefficient on union membership is zero in a wage regression. Consider, for example, the estimates reported in Table 4.1. The coefficient for "Male Union Member" is 0.095 (a wage premium of 9.5%) and the coefficient for "Female Union Member" is 0.022 (a wage premium of 2.2%). These are estimates, not the true values. The question is: Are the true coefficients zero? To answer this question the testing method asks the question: Are the observed estimates compatible with the hypothesis, in the sense that the deviation from the hypothesis can be reasonably explained by stochastic variation? Or are the observed estimates incompatible with the hypothesis, in the sense that that the observed estimates would be highly unlikely if the hypothesis were true?

9.2 Acceptance and Rejection

A hypothesis test either accepts the null hypothesis or rejects the null hypothesis in favor of the alternative hypothesis. We can describe these two decisions as "Accept $\mathbb{H}_0$" and "Reject $\mathbb{H}_0$". In the example given in the previous section the decision is either to accept the hypothesis that union membership does not affect wages, or to reject the hypothesis in favor of the alternative that union membership does affect wages.

The decision is based on the data and so is a mapping from the sample space to the decision set.

This splits the sample space into two regions S_0 and S_1 such that if the observed sample falls into S_0 we accept $\mathbb{H}_0$, while if the sample falls into S_1 we reject $\mathbb{H}_0$. The set S_0 is called the **acceptance region** and the set S_1 the **rejection** or **critical region**.

It is convenient to express this mapping as a real-valued function called a **test statistic**

$$T = T\left(\left(Y_1, X_1\right), ..., \left(Y_n, X_n\right)\right)$$

relative to a **critical value** c. The hypothesis test then consists of the decision rule:

1. Accept $\mathbb{H}_0$ if $T \leq c$.
2. Reject $\mathbb{H}_0$ if $T > c$.

Figure 9.1(b) illustrates the division of the sample space into acceptance and rejection regions.

A test statistic T should be designed so that small values are likely when $\mathbb{H}_0$ is true and large values are likely when $\mathbb{H}_1$ is true. There is a well developed statistical theory concerning the design of optimal tests. We will not review that theory here, but instead refer the reader to Lehmann and Romano (2005). In this chapter we will summarize the main approaches to the design of test statistics.

The most commonly used test statistic is the absolute value of the t-statistic

$$T = |T(\theta_0)| \tag{9.2}$$

where

$$T(\theta) = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} \tag{9.3}$$

is the t-statistic from (7.33), $\widehat{\theta}$ is a point estimator, and $s(\widehat{\theta})$ its standard error. T is an appropriate statistic when testing hypotheses on individual coefficients or real-valued parameters $\theta = h(\beta)$ and θ_0 is the hypothesized value. Quite typically $\theta_0 = 0$, as interest focuses on whether or not a coefficient equals zero, but this is not the only possibility. For example, interest may focus on whether an elasticity θ equals 1, in which case we may wish to test $\mathbb{H}_0 : \theta = 1$.

9.3 Type I Error

A false rejection of the null hypothesis $\mathbb{H}_0$ (rejecting $\mathbb{H}_0$ when $\mathbb{H}_0$ is true) is called a **Type I error**. The probability of a Type I error is called the **size** of the test.

$$\mathbb{P}\left[\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_0 \text{ true}\right] = \mathbb{P}\left[T > c \mid \mathbb{H}_0 \text{ true}\right]. \tag{9.4}$$

The **uniform size** of the test is the supremum of (9.4) across all data distributions which satisfy $\mathbb{H}_0$. A primary goal of test construction is to limit the incidence of Type I error by bounding the size of the test.

For the reasons discussed in Chapter 7, in typical econometric models the exact sampling distributions of estimators and test statistics are unknown and hence we cannot explicitly calculate (9.4). Instead, we typically rely on asymptotic approximations. Suppose that the test statistic has an asymptotic distribution under $\mathbb{H}_0$. That is, when $\mathbb{H}_0$ is true

$$T \underset{d}{\longrightarrow} \xi \tag{9.5}$$

as $n \to \infty$ for some continuously-distributed random variable ξ. This is not a substantive restriction as most conventional econometric tests satisfy (9.5). Let $G(u) = \mathbb{P}\left[\xi \leq u\right]$ denote the distribution of ξ. We call ξ (or G) the **asymptotic null distribution**.

It is desirable to design test statistics T whose asymptotic null distribution G is known and does not depend on unknown parameters. In this case we say that T is **asymptotically pivotal**.

For example, if the test statistic equals the absolute t-statistic from (9.2), then we know from Theorem 7.11 that if $\theta = \theta_0$ (that is, the null hypothesis holds), then $T \xrightarrow[d]{} |Z|$ as $n \to \infty$ where $Z \sim \mathrm{N}(0,1)$. This means that $G(u) = \mathbb{P}[|Z| \le u] = 2\Phi(u) - 1$, the distribution of the absolute value of the standard normal as shown in (7.34). 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:

$$\lim_{n\to\infty} \mathbb{P}[T > c \mid \mathbb{H}_0 \text{ true}] = \mathbb{P}[\xi > c] = 1 - G(c).$$

We see that the asymptotic size of the test is a simple function of the asymptotic null distribution G and the critical value c. For example, the asymptotic size of a test based on the absolute t-statistic with critical value c is $2(1-\Phi(c))$.

In the dominant approach to hypothesis testing the researcher pre-selects a **significance level** $\alpha \in (0,1)$ and then selects c so the asymptotic size is no larger than α. When the asymptotic null distribution G is pivotal we accomplish this by setting c equal to the $1-\alpha$ quantile of the distribution G. (If the distribution G is not pivotal more complicated methods must be used.) We call c the **asymptotic critical value** because it has been selected from the asymptotic null distribution. For example, since $2(1-\Phi(1.96)) = 0.05$ it follows that the 5% asymptotic critical value for the absolute t-statistic is $c = 1.96$. Calculation of normal critical values is done numerically in statistical software. For example, in MATLAB the command is `norminv(1-`α`/2)`.

9.4 t tests

As we mentioned earlier, the most common test of the one-dimensional hypothesis $\mathbb{H}_0 : \theta = \theta_0 \in \mathbb{R}$ against the alternative $\mathbb{H}_1 : \theta \neq \theta_0$ is the absolute value of the t-statistic (9.3). We now formally state its asymptotic null distribution, which is a simple application of Theorem 7.11.

Theorem 9.1 Under Assumptions 7.2, 7.3, and $\mathbb{H}_0 : \theta = \theta_0 \in \mathbb{R}$, $T(\theta_0) \xrightarrow[d]{} Z \sim \mathrm{N}(0,1)$. For c satisfying $\alpha = 2(1-\Phi(c))$, $\mathbb{P}[|T(\theta_0)| > c \mid \mathbb{H}_0] \to \alpha$, and the test "Reject $\mathbb{H}_0$ if $|T(\theta_0)| > c$" has asymptotic size α.

Theorem 9.1 shows that asymptotic critical values can be taken from the normal distribution. As in our discussion of asymptotic confidence intervals (Section 7.13) the critical value could alternatively be taken from the student t distribution, which would be the exact test in the normal regression model (Section 5.12). Indeed, t critical values are the default in packages such as Stata. Since the critical values from the student t distribution are (slightly) larger than those from the normal distribution, student t critical values slightly decrease the rejection probability of the test. In practical applications the difference is typically unimportant unless the sample size is quite small (in which case the asymptotic approximation should be questioned as well).

The alternative hypothesis $\theta \neq \theta_0$ is sometimes called a "two-sided" alternative. In contrast, sometimes we are interested in testing for one-sided alternatives such as $\mathbb{H}_1 : \theta > \theta_0$ or $\mathbb{H}_1 : \theta < \theta_0$. Tests of $\theta = \theta_0$ against $\theta > \theta_0$ or $\theta < \theta_0$ are based on the signed t-statistic $T = T(\theta_0)$. The hypothesis $\theta = \theta_0$ is rejected in favor of $\theta > \theta_0$ if $T > c$ where c satisfies $\alpha = 1 - \Phi(c)$. Negative values of T are not taken as evidence against $\mathbb{H}_0$, as point estimates $\widehat{\theta}$ less than θ_0 do not point to $\theta > \theta_0$. Since the critical values are

taken from the single tail of the normal distribution they are smaller than for two-sided tests. Specifically, the asymptotic 5% critical value is $c = 1.645$. Thus, we reject $\theta = \theta_0$ in favor of $\theta > \theta_0$ if $T > 1.645$.

Conversely, tests of $\theta = \theta_0$ against $\theta < \theta_0$ reject $\mathbb{H}_0$ for negative t-statistics, e.g. if $T < -c$. Large positive values of T are not evidence for $\mathbb{H}_1 : \theta < \theta_0$. An asymptotic 5% test rejects if $T < -1.645$.

There seems to be an ambiguity. Should we use the two-sided critical value 1.96 or the one-sided critical value 1.645? The answer is that in most cases the two-sided critical value is appropriate. We should use the one-sided critical values only when the parameter space is known to satisfy a one-sided restriction such as $\theta \geq \theta_0$. This is when the test of $\theta = \theta_0$ against $\theta > \theta_0$ makes sense. If the restriction $\theta \geq \theta_0$ is not known *a priori* then imposing this restriction to test $\theta = \theta_0$ against $\theta > \theta_0$ does not makes sense. Since linear regression coefficients typically do not have *a priori* sign restrictions, the standard convention is to use two-sided critical values.

This may seem contrary to the way testing is presented in statistical textbooks which often focus on one-sided alternative hypotheses. The latter focus is primarily for pedagogy as the one-sided theoretical problem is cleaner and easier to understand.

9.5 Type II Error and Power

A false acceptance of the null hypothesis $\mathbb{H}_0$ (accepting $\mathbb{H}_0$ when $\mathbb{H}_1$ is true) is called a **Type II error**. The rejection probability under the alternative hypothesis is called the **power** of the test, and equals 1 minus the probability of a Type II error:

$$\pi(\theta) = \mathbb{P}\left[\text{Reject } \mathbb{H}_0 \mid \mathbb{H}_1 \text{ true}\right] = \mathbb{P}\left[T > c \mid \mathbb{H}_1 \text{ true}\right].$$

We call $\pi(\theta)$ the **power function** and is written as a function of θ to indicate its dependence on the true value of the parameter θ.

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 θ, and for a well-behaved test the power is increasing both as θ moves away from the null hypothesis θ_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 Table 9.1.

Table 9.1: Hypothesis Testing Decisions

	Accept $\mathbb{H}_0$	Reject $\mathbb{H}_0$
$\mathbb{H}_0$ true	Correct Decision	Type I Error
$\mathbb{H}_1$ true	Type II Error	Correct Decision

Given a test statistic T, increasing the critical value c increases the acceptance region S_0 while decreasing the rejection region S_1. This decreases the likelihood of a Type I error (decreases the size) but increases the likelihood of a Type II error (decreases the power). Thus the choice of c involves a trade-off between size and the power. This is why the significance level α of the test cannot be set arbitrarily small. Otherwise the test will not have meaningful power.

It is important to consider the power of a test when interpreting hypothesis tests as an overly narrow focus on size can lead to poor decisions. For example, it is easy to design a test which has perfect size yet has trivial power. Specifically, for any hypothesis we can use the following test: Generate a random variable $U \sim U[0,1]$ and reject $\mathbb{H}_0$ if $U < \alpha$. This test has exact size of α. Yet the test also has power precisely equal to α. When the power of a test equals the size we say that the test has **trivial power**. Nothing is learned from such a test.

9.6 Statistical Significance

Testing requires a pre-selected choice of significance level α yet there is no objective scientific basis for choice of α. Nevertheless, the common practice is to set $\alpha = 0.05$ (5%). Alternative common 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 5% critical value is to ensure that Type I errors should be relatively unlikely – that the decision "Reject $\mathbb{H}_0$" has scientific strength – yet the test retains power against reasonable alternatives. The decision "Reject $\mathbb{H}_0$" means that the evidence is inconsistent with the null hypothesis in the sense that it is relatively unlikely (1 in 20) that data generated by the null hypothesis would yield the observed test result.

In contrast, the decision "Accept $\mathbb{H}_0$" is not a strong statement. It does not mean that the evidence supports $\mathbb{H}_0$, only that there is insufficient evidence to reject $\mathbb{H}_0$. Because of this it is more accurate to use the label "Do not Reject $\mathbb{H}_0$" instead of "Accept $\mathbb{H}_0$".

When a test rejects $\mathbb{H}_0$ at the 5% significance level it is common to say that the statistic is **statistically significant** and if the test accepts $\mathbb{H}_0$ it is common to say that the statistic is **not statistically significant** or that it is **statistically insignificant**. It is helpful to remember that this is simply a compact way of saying "Using the statistic T the hypothesis $\mathbb{H}_0$ can [cannot] be rejected at the asymptotic 5% level." Furthermore, when the null hypothesis $\mathbb{H}_0 : \theta = 0$ is rejected it is common to say that the coefficient θ is statistically significant, because the test has rejected the hypothesis that the coefficient is equal to zero.

Let us return to the example about the union wage premium as measured in Table 4.1. The absolute t-statistic for the coefficient on "Male Union Member" is $0.095/0.020 = 4.7$, which is greater than the 5% asymptotic critical value of 1.96. Therefore we reject the hypothesis that union membership does not affect wages for men. In this case we can say that union membership is statistically significant for men. However, the absolute t-statistic for the coefficient on "Female Union Member" is $0.023/0.020 = 1.2$, which is less than 1.96 and therefore we do not reject the hypothesis that union membership does not affect wages for women. In this case we find that membership for women is not statistically significant.

When a test accepts a null hypothesis (when a test is not statistically significant) a common misinterpretation is that this is evidence that the null hypothesis is true. This is incorrect. Failure to reject is by itself not evidence. Without an analysis of power we do not know the likelihood of making a Type II error and thus are uncertain. In our wage example it would be a mistake to write that "the regression finds that female union membership has no effect on wages". This is an incorrect and most unfortunate interpretation. The test has failed to reject the hypothesis that the coefficient is zero but that does not mean that the coefficient is actually zero.

When a test rejects a null hypothesis (when a test is statistically significant) it is strong evidence against the hypothesis (because if the hypothesis were true then rejection is an unlikely event). Rejection should be taken as evidence against the null hypothesis. However, we can never conclude that the null hypothesis is indeed false as we cannot exclude the possibility that we are making a Type I error.

Perhaps more importantly, there is an important distinction between statistical and economic significance. If we correctly reject the hypothesis $\mathbb{H}_0 : \theta = 0$ it means that the true value of θ is non-zero. This includes the possibility that θ may be non-zero but close to zero in magnitude. This only makes sense if we interpret the parameters in the context of their relevant models. In our wage regression example we might consider wage effects of 1% magnitude or less as being "close to zero". In a log wage regression this corresponds to a dummy variable with a coefficient less than 0.01. If the standard error is sufficiently small (less than 0.005) then a coefficient estimate of 0.01 will be statistically significant but not economically significant. This occurs frequently in applications with very large sample sizes where standard errors can be quite small.

The solution is to focus whenever possible on confidence intervals and the economic meaning of the

coefficients. For example, if the coefficient estimate is 0.005 with a standard error of 0.002 then a 95% confidence interval would be [0.001, 0.009] indicating that the true effect is likely between 0% and 1%, and hence is slightly positive but small. This is much more informative than the misleading statement "the effect is statistically positive".

9.7 P-Values

Continuing with the wage regression estimates reported in Table 4.1, consider another question: Does marriage status affect wages? To test the hypothesis that marriage status has no effect on wages, we examine the t-statistics for the coefficients on "Married Male" and "Married Female" in Table 4.1, which are $0.211/0.010 = 22$ and $0.016/0.010 = 1.7$, respectively. The first exceeds the asymptotic 5% critical value of 1.96 so we reject the hypothesis for men. The second is smaller than 1.96 so we fail to reject the hypothesis for women. Taking a second look at the statistics we see that the statistic for men (22) is exceptionally high and that for women (1.7) is only slightly below the critical value. Suppose that the t-statistic for women were slightly increased to 2.0. This is larger than the critical value so would lead to the decision "Reject $\mathbb{H}_0$" rather than "Accept $\mathbb{H}_0$". Should we really be making a different decision if the t-statistic is 2.0 rather than 1.7? The difference in values is small, shouldn't the difference in the decision be also small? Thinking through these examples it seems unsatisfactory to simply report "Accept $\mathbb{H}_0$" or "Reject $\mathbb{H}_0$". These two decisions do not summarize the evidence. Instead, the magnitude of the statistic T suggests a "degree of evidence" against $\mathbb{H}_0$. How can we take this into account?

The answer is to report what is known as the **asymptotic p-value**

$$p = 1 - G(T).$$

Since the distribution function G is monotonically increasing, the p-value is a monotonically decreasing function of T and is an equivalent test statistic. Instead of rejecting $\mathbb{H}_0$ at the significance level α if $T > c$, we can reject $\mathbb{H}_0$ if $p < \alpha$. Thus it is sufficient to report p, and let the reader decide. In practice, the p-value is calculated numerically. For example, in MATLAB the command is `2*(1-normalcdf(abs(t)))`.

It is instructive to interpret p as the **marginal significance level**: the smallest value of α for which the test T "rejects" the null hypothesis. That is, $p = 0.11$ means that T rejects $\mathbb{H}_0$ for all significance levels greater than 0.11, but fails to reject $\mathbb{H}_0$ for significance levels less than 0.11.

Furthermore, the asymptotic p-value has a very convenient asymptotic null distribution. Since $T \underset{d}{\longrightarrow} \xi$ under $\mathbb{H}_0$, then $p = 1 - G(T) \underset{d}{\longrightarrow} 1 - G(\xi)$, which has the distribution

$$\begin{aligned}
\mathbb{P}[1 - G(\xi) \leq u] &= \mathbb{P}[1 - u \leq G(\xi)] \\
&= 1 - \mathbb{P}\left[\xi \leq G^{-1}(1-u)\right] \\
&= 1 - G\left(G^{-1}(1-u)\right) \\
&= 1 - (1-u) \\
&= u,
\end{aligned}$$

which is the uniform distribution on $[0,1]$. (This calculation assumes that $G(u)$ is strictly increasing which is true for conventional asymptotic distributions such as the normal.) Thus $p \underset{d}{\longrightarrow} \mathrm{U}[0,1]$. This means that the "unusualness" of p is easier to interpret than the "unusualness" of T.

An important caveat is that the p-value p should not be interpreted as the probability that either hypothesis is true. A common mis-interpretation is that p is the probability "that the null hypothesis is true." This is incorrect. Rather, p is the marginal significance level – a measure of the strength of information against the null hypothesis.

For a t-statistic the p-value can be calculated either using the normal distribution or the student t distribution, the latter presented in Section 5.12. p-values calculated using the student t will be slightly larger, though the difference is small when the sample size is large.

Returning to our empirical example, for the test that the coefficient on "Married Male" is zero the p-value is 0.000. This means that it would be nearly impossible to observe a t-statistic as large as 22 when the true value of the coefficient is zero. When presented with such evidence we can say that we "strongly reject" the null hypothesis, that the test is "highly significant", or that "the test rejects at any conventional critical value". In contrast, the p-value for the coefficient on "Married Female" is 0.094. In this context it is typical to say that the test is "close to significant", meaning that the p-value is larger than 0.05, but not too much larger.

A related but 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, and append a triple asterisk (***) for a test which exceeds the 1% critical value. Such a practice can be better than a table of raw test statistics as the asterisks permit a quick interpretation of significance. On the other hand, asterisks are inferior to p-values, which are also easy and quick to interpret. The goal is essentially the same; it is wiser to report p-values whenever possible and avoid the use of asterisks.

Our recommendation is that the best empirical practice is to compute and report the asymptotic p-value p rather than simply the test statistic T, the binary decision Accept/Reject, or appending asterisks. The p-value is a simple statistic, easy to interpret, and contains more information than the other choices.

We now summarize the main features of hypothesis testing.

1. Select a significance level α.
2. Select a test statistic T with asymptotic distribution $T \underset{d}{\longrightarrow} \xi$ under $\mathbb{H}_0$.
3. Set the asymptotic critical value c so that $1 - G(c) = \alpha$, where G is the distribution function of ξ.
4. Calculate the asymptotic p-value $p = 1 - G(T)$.
5. Reject $\mathbb{H}_0$ if $T > c$, or equivalently $p < \alpha$.
6. Accept $\mathbb{H}_0$ if $T \leq c$, or equivalently $p \geq \alpha$.
7. Report p to summarize the evidence concerning $\mathbb{H}_0$ versus $\mathbb{H}_1$.

9.8 t-ratios and the Abuse of Testing

In Section 4.19 we argued that a good applied practice is to report coefficient estimates $\widehat{\theta}$ and standard errors $s(\widehat{\theta})$ for all coefficients of interest in estimated models. With $\widehat{\theta}$ and $s(\widehat{\theta})$ the reader can easily construct confidence intervals $[\widehat{\theta} \pm 2s(\widehat{\theta})]$ and t-statistics $\left(\widehat{\theta} - \theta_0\right) / s(\widehat{\theta})$ for hypotheses of interest.

Some applied papers (especially older ones) report t-ratios $T = \widehat{\theta} / s(\widehat{\theta})$ instead of standard errors. This is poor econometric practice. While the same information is being reported (you can back out standard errors by division, e.g. $s(\widehat{\theta}) = \widehat{\theta} / T$), standard errors are generally more helpful to readers than t-ratios. Standard errors help the reader focus on the estimation precision and confidence intervals, while t-ratios focus attention on statistical significance. While statistical significance is important, it is less important that the parameter estimates themselves and their confidence intervals. The focus should be on the meaning of the parameter estimates, their magnitudes, and their interpretation, not on

listing which variables have significant (e.g. non-zero) coefficients. In many modern applications sample sizes are very large so standard errors can be very small. Consequently t-ratios can be large even if the coefficient estimates are economically small. In such contexts it may not be interesting to announce "The coefficient is non-zero!" Instead, what is interesting to announce is that "The coefficient estimate is economically interesting!"

In particular, some applied papers report coefficient estimates and t-ratios and limit their discussion of the results to describing which variables are "significant" (meaning that their t-ratios exceed 2) and the signs of the coefficient estimates. This is very poor empirical work and should be studiously avoided. It is also a recipe for banishment of your work to lower tier economics journals.

Fundamentally, the common t-ratio is a test for the hypothesis that a coefficient equals zero. This should be reported and discussed when this is an interesting economic hypothesis of interest. But if this is not the case it is distracting.

One problem is that standard packages, such as Stata, by default report t-statistics and p-values for every estimated coefficient. While this can be useful (as a user doesn't need to explicitly ask to test a desired coefficient) it can be misleading as it may unintentionally suggest that the entire list of t-statistics and p-values are important. Instead, a user should focus on tests of scientifically motivated hypotheses.

In general, when a coefficient θ is of interest it is constructive to focus on the point estimate, its standard error, and its confidence interval. The point estimate gives our "best guess" for the value. The standard error is a measure of precision. The confidence interval gives us the range of values consistent with the data. If the standard error is large then the point estimate is not a good summary about θ. The endpoints of the confidence interval describe the bounds on the likely possibilities. If the confidence interval embraces too broad a set of values for θ then the dataset is not sufficiently informative to render useful inferences about θ. On the other hand if the confidence interval is tight then the data have produced an accurate estimate and the focus should be on the value and interpretation of this estimate. In contrast, the statement "the t-ratio is highly significant" has little interpretive value.

The above discussion requires that the researcher knows what the coefficient θ means (in terms of the economic problem) and can interpret values and magnitudes, not just signs. This is critical for good applied econometric practice.

For example, consider the question about the effect of marriage status on mean log wages. We had found that the effect is "highly significant" for men and "close to significant" for women. Now, let's construct asymptotic 95% confidence intervals for the coefficients. The one for men is [0.19, 0.23] and that for women is [−0.00, 0.03]. This shows that average wages for married men are about 19-23% higher than for unmarried men, which is substantial, while the difference for women is about 0-3%, which is small. These *magnitudes* are more informative than the results of the hypothesis tests.

9.9 Wald Tests

The t-test is appropriate when the null hypothesis is a real-valued restriction. More generally there may be multiple restrictions on the coefficient vector β. Suppose that we have $q > 1$ restrictions which can be written in the form (9.1). It is natural to estimate $\theta = r(\beta)$ by the plug-in estimator $\widehat{\theta} = r\left(\widehat{\beta}\right)$. To test $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ one approach is to measure the magnitude of the discrepancy $\widehat{\theta} - \theta_0$. As this is a vector there is more than one measure of its length. One simple measure is the weighted quadratic form known as the **Wald statistic**. This is (7.37) evaluated at the null hypothesis

$$W = W(\theta_0) = \left(\widehat{\theta} - \theta_0\right)' \widehat{\boldsymbol{V}}_{\widehat{\theta}}^{-1} \left(\widehat{\theta} - \theta_0\right) \tag{9.6}$$

where $\widehat{\boldsymbol{V}}_{\widehat{\theta}} = \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\widehat{\beta}}\widehat{\boldsymbol{R}}$ is an estimator of $\boldsymbol{V}_{\widehat{\theta}}$ and $\widehat{\boldsymbol{R}} = \dfrac{\partial}{\partial\beta} r(\widehat{\beta})'$. Notice that we can write W alternatively as

$$W = n\left(\widehat{\theta} - \theta_0\right)' \widehat{\boldsymbol{V}}_{\theta}^{-1} \left(\widehat{\theta} - \theta_0\right)$$

using the asymptotic variance estimator $\widehat{\boldsymbol{V}}_{\theta}$, or we can write it directly as a function of $\widehat{\beta}$ as

$$W = \left(r(\widehat{\beta}) - \theta_0\right)' \left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\widehat{\beta}}\widehat{\boldsymbol{R}}\right)^{-1} \left(r(\widehat{\beta}) - \theta_0\right).$$

Also, when $r(\beta) = \boldsymbol{R}'\beta$ is a linear function of β, then the Wald statistic simplifies to

$$W = \left(\boldsymbol{R}'\widehat{\beta} - \theta_0\right)' \left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\widehat{\beta}}\boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\beta} - \theta_0\right).$$

The Wald statistic W is a weighted Euclidean measure of the length of the vector $\widehat{\theta} - \theta_0$. When $q = 1$ then $W = T^2$, the square of the t-statistic, so hypothesis tests based on W and $|T|$ are equivalent. The Wald statistic (9.6) is a generalization of the t-statistic to the case of multiple restrictions. As the Wald statistic is symmetric in the argument $\widehat{\theta} - \theta_0$ it treats positive and negative alternatives symmetrically. Thus the inherent alternative is always two-sided.

As shown in Theorem 7.13, when β satisfies $r(\beta) = \theta_0$ then $W \xrightarrow[d]{} \chi^2_q$, a chi-square random variable with q degrees of freedom. Let $G_q(u)$ denote the χ^2_q distribution function. For a given significance level α the asymptotic critical value c satisfies $\alpha = 1 - G_q(c)$. For example, the 5% critical values for $q = 1$, $q = 2$, and $q = 3$ are 3.84, 5.99, and 7.82, respectively, and in general the level α critical value can be calculated in MATLAB as `chi2inv(1-α,q)`. An asymptotic test rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $W > c$. As with t-tests, it is conventional to describe a Wald test as "significant" if W exceeds the 5% asymptotic critical value.

Theorem 9.2 Under Assumptions 7.2, 7.3, 7.4, and $\mathbb{H}_0 : \theta = \theta_0 \in \mathbb{R}^q$, then $W \xrightarrow[d]{} \chi^2_q$. For c satisfying $\alpha = 1 - G_q(c)$, $\mathbb{P}(W > c \mid \mathbb{H}_0) \longrightarrow \alpha$ so the test "Reject $\mathbb{H}_0$ if $W > c$" has asymptotic size α.

Notice that the asymptotic distribution in Theorem 9.2 depends solely on q, the number of restrictions being tested. It does not depend on k, the number of parameters estimated.

The asymptotic p-value for W is $p = 1 - G_q(W)$, and this is particularly useful when testing multiple restrictions. For example, if you write that a Wald test on eight restrictions ($q = 8$) has the value $W = 11.2$ it is difficult for a reader to assess the magnitude of this statistic unless they have quick access to a statistical table or software. Instead, if you write that the p-value is $p = 0.19$ (as is the case for $W = 11.2$ and $q = 8$) then it is simple for a reader to interpret its magnitude as "insignificant". To calculate the asymptotic p-value for a Wald statistic in MATLAB use the command `1-chi2cdf(w,q)`.

Some packages (including Stata) and papers report F versions of Wald statistics. For any Wald statistic W which tests a q-dimensional restriction, the F version of the test is

$$F = W/q.$$

When F is reported, it is conventional to use $F_{q,n-k}$ critical values and p-values rather than χ^2_q values. The connection between Wald and F statistics is demonstrated in Section 9.14 where we show that when Wald statistics are calculated using a homoskedastic covariance matrix then $F = W/q$ is identicial to

the F statistic of (5.19). While there is no formal justification to using the $F_{q,n-k}$ distribution for non-homoskedastic covariance matrices, the $F_{q,n-k}$ distribution provides continuity with the exact distribution theory under normality and is a bit more conservative than the χ^2_q distribution. (Furthermore, the difference is small when $n-k$ is moderately large.)

To implement a test of zero restrictions in Stata an easy method is to use the command `test X1 X2` where `X1` and `X2` are the names of the variables whose coefficients are hypothesized to equal zero. The F version of the Wald statistic is reported using the covariance matrix calculated by the method specified in the regression command. A p-value is reported, calculated using the $F_{q,n-k}$ distribution.

To illustrate, consider the empirical results presented in Table 4.1. The hypothesis "Union membership does not affect wages" is the joint restriction that both coefficients on "Male Union Member" and "Female Union Member" are zero. We calculate the Wald statistic for this joint hypothesis and find $W = 23$ (or $F = 12.5$) with a p-value of $p = 0.000$. Thus we reject the null hypothesis in favor of the alternative that at least one of the coefficients is non-zero. This does not mean that both coefficients are non-zero, just that one of the two is non-zero. Therefore examining both the joint Wald statistic and the individual t-statistics is useful for interpretation.

As a second example from the same regression, take the hypothesis that married status has no effect on mean wages for women. This is the joint restriction that the coefficients on "Married Female" and "Formerly Married Female" are zero. The Wald statistic for this hypothesis is $W = 6.4$ ($F = 3.2$) with a p-value of 0.04. Such a p-value is typically called "marginally significant" in the sense that it is slightly smaller than 0.05.

The Wald statistic was proposed by Wald (1943).

Abraham Wald

The Hungarian mathematician/statistician/econometrician Abraham Wald (1902-1950) developed an optimality property for the Wald test in terms of weighted average power. He also developed the field of sequential testing, the design of experiments, and one of the first instrumental variable estimators.

9.10 Homoskedastic Wald Tests

If the error is known to be homoskedastic then it is appropriate to use the homoskedastic Wald statistic (7.38) which replaces $\widehat{\boldsymbol{V}}_{\widehat{\theta}}$ with the homoskedastic estimator $\widehat{\boldsymbol{V}}^0_{\widehat{\theta}}$. This statistic equals

$$\begin{aligned} W^0 &= \left(\widehat{\theta}-\theta_0\right)' \left(\widehat{\boldsymbol{V}}^0_{\widehat{\theta}}\right)^{-1} \left(\widehat{\theta}-\theta_0\right) \\ &= \left(r(\widehat{\beta})-\theta_0\right)' \left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\widehat{\boldsymbol{R}}\right)^{-1} \left(r(\widehat{\beta})-\theta_0\right)/s^2. \end{aligned}$$

In the case of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\beta = \theta_0$ we can write this as

$$W^0 = \left(\boldsymbol{R}'\widehat{\beta}-\theta_0\right)' \left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1} \left(\boldsymbol{R}'\widehat{\beta}-\theta_0\right)/s^2. \tag{9.7}$$

We call W^0 a **homoskedastic Wald statistic** as it is appropriate when the errors are conditionally homoskedastic.

When $q = 1$ then $W^0 = T^2$, the square of the t-statistic where the latter is computed with a homoskedastic standard error.

Theorem 9.3 Under Assumptions 7.2 and 7.3, $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2 > 0$, and $\mathbb{H}_0 : \theta = \theta_0 \in \mathbb{R}^q$, then $W^0 \underset{d}{\longrightarrow} \chi^2_q$. For c satisfying $\alpha = 1 - G_q(c)$, $\mathbb{P}\left[W^0 > c \mid \mathbb{H}_0\right] \longrightarrow \alpha$ so the test "Reject $\mathbb{H}_0$ if $W^0 > c$" has asymptotic size α.

9.11 Criterion-Based Tests

The Wald statistic is based on the length of the vector $\widehat{\theta} - \theta_0$: the discrepancy between the estimator $\widehat{\theta} = r(\widehat{\beta})$ and the hypothesized value θ_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 $J(\beta)$ with $\beta \in B$, which is minimized for estimation, and the goal is to test $\mathbb{H}_0 : \beta \in B_0$ versus $\mathbb{H}_1 : \beta \notin B_0$ where $B_0 \subset \beta$. Minimizing the criterion function over B and B_0 we obtain the unrestricted and restricted estimators

$$\widehat{\beta} = \underset{\beta \in B}{\operatorname{argmin}}\ J\left(\beta\right)$$
$$\widetilde{\beta} = \underset{\beta \in B_0}{\operatorname{argmin}}\ J\left(\beta\right).$$

The **criterion-based statistic** for $\mathbb{H}_0$ versus $\mathbb{H}_1$ is proportional to

$$J = \min_{\beta \in B_0} J\left(\beta\right) - \min_{\beta \in B} J\left(\beta\right) = J(\widetilde{\beta}) - J(\widehat{\beta}).$$

The criterion-based statistic J is sometimes called a **distance** statistic, a **minimum-distance** statistic, or a **likelihood-ratio-like** statistic.

Since B_0 is a subset of B, $J(\widetilde{\beta}) \geq J(\widehat{\beta})$ and thus $J \geq 0$. The statistic J measures the cost on the criterion of imposing the null restriction $\beta \in B_0$.

9.12 Minimum Distance Tests

The minimum distance test is based on the minimum distance criterion (8.19)

$$J\left(\beta\right) = n\left(\widehat{\beta} - \beta\right)' \widehat{\boldsymbol{W}} \left(\widehat{\beta} - \beta\right) \tag{9.8}$$

with $\widehat{\beta}$ the unrestricted least squares estimator. The restricted estimator $\widetilde{\beta}_{\text{md}}$ minimizes (9.8) subject to $\beta \in B_0$. Observing that $J(\widehat{\beta}) = 0$, the minimum distance statistic simplifies to

$$J = J(\widetilde{\beta}_{\text{md}}) = n\left(\widehat{\beta} - \widetilde{\beta}_{\text{md}}\right)' \widehat{\boldsymbol{W}} \left(\widehat{\beta} - \widetilde{\beta}_{\text{md}}\right). \tag{9.9}$$

The efficient minimum distance estimator $\widetilde{\beta}_{\text{emd}}$ is obtained by setting $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{V}}_\beta^{-1}$ in (9.8) and (9.9). The efficient minimum distance statistic for $\mathbb{H}_0 : \beta \in B_0$ is therefore

$$J^* = n\left(\widehat{\beta} - \widetilde{\beta}_{\text{emd}}\right)' \widehat{\boldsymbol{V}}_\beta^{-1} \left(\widehat{\beta} - \widetilde{\beta}_{\text{emd}}\right). \tag{9.10}$$

Consider the class of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\beta = \theta_0$. In this case we know from (8.25) that the efficient minimum distance estimator $\widetilde{\beta}_{\text{emd}}$ subject to the constraint $\boldsymbol{R}'\beta = \theta_0$ is

$$\widetilde{\beta}_{\text{emd}} = \widehat{\beta} - \widehat{\boldsymbol{V}}_\beta \boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta \boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \theta_0\right)$$

and thus

$$\widehat{\beta} - \widetilde{\beta}_{\text{emd}} = \widehat{\boldsymbol{V}}_\beta \boldsymbol{R} \left(\boldsymbol{R}' \widehat{\boldsymbol{V}}_\beta \boldsymbol{R} \right)^{-1} \left(\boldsymbol{R}' \widehat{\beta} - \theta_0 \right).$$

Substituting into (9.10) we find

$$\begin{aligned} J^* &= n \left(\boldsymbol{R}' \widehat{\beta} - \theta_0 \right)' \left(\boldsymbol{R}' \widehat{\boldsymbol{V}}_\beta \boldsymbol{R} \right)^{-1} \boldsymbol{R}' \widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{V}}_{\boldsymbol{\beta}}^{-1} \widehat{\boldsymbol{V}}_\beta \boldsymbol{R} \left(\boldsymbol{R}' \widehat{\boldsymbol{V}}_\beta \boldsymbol{R} \right)^{-1} \left(\boldsymbol{R}' \widehat{\beta} - \theta_0 \right) \\ &= n \left(\boldsymbol{R}' \widehat{\beta} - \theta_0 \right)' \left(\boldsymbol{R}' \widehat{\boldsymbol{V}}_\beta \boldsymbol{R} \right)^{-1} \left(\boldsymbol{R}' \widehat{\beta} - \theta_0 \right) \\ &= W, \end{aligned}$$

which is the Wald statistic (9.6).

Thus for linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\beta = \theta_0$, the efficient minimum distance statistic J^* is identical to the Wald statistic (9.6). For nonlinear hypotheses, however, the Wald and minimum distance statistics are different.

Newey and West (1987a) established the asymptotic null distribution of J^*.

Theorem 9.4 Under Assumptions 7.2, 7.3, 7.4, and $\mathbb{H}_0 : \theta = \theta_0 \in \mathbb{R}^q$, $J^* \underset{d}{\longrightarrow} \chi^2_q$.

Testing using the minimum distance statistic J^* is similar to testing using the Wald statistic W. Critical values and p-values are computed using the χ^2_q distribution. $\mathbb{H}_0$ is rejected in favor of $\mathbb{H}_1$ if J^* exceeds the level α critical value, which can be calculated in MATLAB as `chi2inv(1-α,q)`. The asymptotic p-value is $p = 1 - G_q(J^*)$. In MATLAB, use the command `1-chi2cdf(J,q)`.

We now demonstrate Theorem 9.4. The conditions of Theorem 8.10 hold, because $\mathbb{H}_0$ implies Assumption 8.1. From (8.54) with $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{V}}_\beta$, we see that

$$\begin{aligned} \sqrt{n} \left(\widehat{\beta} - \widetilde{\beta}_{\text{emd}} \right) &= \widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}} \left(\boldsymbol{R}_n^{*\prime} \widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}} \right)^{-1} \boldsymbol{R}_n^{*\prime} \sqrt{n} \left(\widehat{\beta} - \beta \right) \\ &\underset{d}{\longrightarrow} \boldsymbol{V}_\beta \boldsymbol{R} \left(\boldsymbol{R}' \boldsymbol{V}_\beta \boldsymbol{R} \right)^{-1} \boldsymbol{R}' \mathrm{N}(0, \boldsymbol{V}_\beta) = \boldsymbol{V}_\beta \boldsymbol{R}\, Z \end{aligned}$$

where $Z \sim \mathrm{N}(0, \left(\boldsymbol{R}' \boldsymbol{V}_\beta \boldsymbol{R} \right)^{-1})$. Thus

$$J^* = n \left(\widehat{\beta} - \widetilde{\beta}_{\text{emd}} \right)' \widehat{\boldsymbol{V}}_\beta^{-1} \left(\widehat{\beta} - \widetilde{\beta}_{\text{emd}} \right) \underset{d}{\longrightarrow} Z' \boldsymbol{R}' \boldsymbol{V}_\beta \boldsymbol{V}_\beta^{-1} \boldsymbol{V}_\beta \boldsymbol{R}\, Z = Z' \left(\boldsymbol{R}' \boldsymbol{V}_\beta \boldsymbol{R} \right) Z = \chi^2_q$$

as claimed.

9.13 Minimum Distance Tests Under Homoskedasticity

If we set $\widehat{\boldsymbol{W}} = \widehat{\boldsymbol{Q}}_{XX} / s^2$ in (9.8) we obtain the criterion (8.20)

$$J^0(\beta) = n \left(\widehat{\beta} - \beta \right)' \widehat{\boldsymbol{Q}}_{XX} \left(\widehat{\beta} - \beta \right) / s^2.$$

A minimum distance statistic for $\mathbb{H}_0 : \beta \in B_0$ is

$$J^0 = \min_{\beta \in B_0} J^0(\beta).$$

Equation (8.21) showed that $\text{SSE}(\beta) = n\widehat{\sigma}^2 + s^2 J^0(\beta)$. So the minimizers of $\text{SSE}(\beta)$ and $J^0(\beta)$ are identical. Thus the constrained minimizer of $J^0(\beta)$ is constrained least squares

$$\widetilde{\beta}_{\text{cls}} = \underset{\beta \in B_0}{\operatorname{argmin}} J^0(\beta) = \underset{\beta \in B_0}{\operatorname{argmin}} \text{SSE}(\beta) \tag{9.11}$$

and therefore

$$J_n^0 = J_n^0(\widetilde{\beta}_{\text{cls}}) = n\left(\widehat{\beta} - \widetilde{\beta}_{\text{cls}}\right)' \widehat{\boldsymbol{Q}}_{XX}\left(\widehat{\beta} - \widetilde{\beta}_{\text{cls}}\right) / s^2.$$

In the special case of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\beta = \theta_0$, the constrained least squares estimator subject to $\boldsymbol{R}'\beta = \theta_0$ has the solution (8.9)

$$\widetilde{\beta}_{\text{cls}} = \widehat{\beta} - \widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \theta_0\right)$$

and solving we find

$$J^0 = n\left(\boldsymbol{R}'\widehat{\beta} - \theta_0\right)'\left(\boldsymbol{R}'\widehat{\boldsymbol{Q}}_{XX}^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \theta_0\right) / s^2 = W^0.$$

This is the homoskedastic Wald statistic (9.7). Thus for testing linear hypotheses, homoskedastic minimum distance and Wald statistics agree.

For nonlinear hypotheses they disagree, but have the same null asymptotic distribution.

Theorem 9.5 Under Assumptions 7.2 and 7.3, $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2 > 0$, and $\mathbb{H}_0 : \theta = \theta_0 \in \mathbb{R}^q$, then $J^0 \xrightarrow[d]{} \chi_q^2$.

9.14 F Tests

In Section 5.13 we introduced the F test for exclusion restrictions in the normal regression model. In this section we generalize this test to a broader set of restrictions. Let $B_0 \subset \mathbb{R}^k$ be a constrained parameter space which imposes q restrictions on β.

Let $\widehat{\beta}_{\text{ols}}$ be the unrestricted least squares estimator and let $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \left(Y_i - X_i'\widehat{\beta}_{\text{ols}}\right)^2$ be the associated estimator of σ^2. Let $\widetilde{\beta}_{\text{cls}}$ be the CLS estimator (9.11) satisfying $\widetilde{\beta}_{\text{cls}} \in B_0$ and let $\widetilde{\sigma}^2 = n^{-1}\sum_{i=1}^n \left(Y_i - X_i'\widetilde{\beta}_{\text{cls}}\right)^2$ be the associated estimator of σ^2. The F statistic for testing $\mathbb{H}_0 : \beta \in B_0$ is

$$F = \frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right) / q}{\widehat{\sigma}^2 / (n-k)}. \tag{9.12}$$

We can alternatively write

$$F = \frac{\text{SSE}(\widetilde{\beta}_{\text{cls}}) - \text{SSE}(\widehat{\beta}_{\text{ols}})}{q s^2} \tag{9.13}$$

where $\text{SSE}(\beta) = \sum_{i=1}^n \left(Y_i - X_i'\beta\right)^2$ is the sum-of-squared errors.

This shows that F is a criterion-based statistic. Using (8.21) we can also write $F = J^0 / q$, so the F statistic is identical to the homoskedastic minimum distance statistic divided by the number of restrictions q.

As we discussed in the previous section, in the special case of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\beta = \theta_0$, $J^0 = W^0$. It follows that in this case $F = W^0 / q$. Thus for linear restrictions the F statistic equals the homoskedastic Wald statistic divided by q. It follows that they are equivalent tests for $\mathbb{H}_0$ against $\mathbb{H}_1$.

Theorem 9.6 For tests of linear hypotheses $\mathbb{H}_0 : \boldsymbol{R}'\beta = \theta_0 \in \mathbb{R}^q$, the F statistic equals $F = W^0/q$ where W^0 is the homoskedastic Wald statistic. Thus under 7.2, $\mathbb{E}\left[e^2 \mid X\right] = \sigma^2 > 0$, and $\mathbb{H}_0 : \theta = \theta_0$, then $F \xrightarrow[d]{} \chi^2_q/q$.

When using an F statistic it is conventional to use the $F_{q,n-k}$ distribution for critical values and p-values. Critical values are given in MATLAB by `finv(1-α,q,n-k)` and p-values by `1-fcdf(F,q,n-k)`. Alternatively, the χ^2_q/q distribution can be used, using `chi2inv(1-α,q)/q` and `1-chi2cdf(F*q,q)`, respectively. Using the $F_{q,n-k}$ distribution is a prudent small sample adjustment which yields exact answers if the errors are normal and otherwise slightly increasing the critical values and p-values relative to the asymptotic approximation. Once again, if the sample size is small enough that the choice makes a difference then probably we shouldn't be trusting the asymptotic approximation anyway!

An elegant feature about (9.12) or (9.13) is that they are directly computable from the standard output from two simple OLS regressions, as the sum of squared errors (or regression variance) is a typical printed output from statistical packages and is often reported in applied tables. Thus F can be calculated by hand from standard reported statistics even if you don't have the original data (or if you are sitting in a seminar and listening to a presentation!).

If you are presented with an F statistic (or a Wald statistic, as you can just divide by q) but don't have access to critical values, a useful rule of thumb is to know that for large n the 5% asymptotic critical value is decreasing as q increases and is less than 2 for $q \geq 7$.

A word of warning: In many statistical packages when an OLS regression is estimated an "F-statistic" is automatically reported even though no hypothesis test was requested. What the package is reporting is an F statistic of the hypothesis that all slope coefficients[1] are zero. This was a popular statistic in the early days of econometric reporting when sample sizes were very small and researchers wanted to know if there was "any explanatory power" to their regression. This is rarely an issue today as sample sizes are typically sufficiently large that this F statistic is nearly always highly significant. While there are special cases where this F statistic is useful these cases are not typical. As a general rule there is no reason to report this F statistic.

9.15 Hausman Tests

Hausman (1978) introduced a general idea about how to test a hypothesis $\mathbb{H}_0$. If you have two estimators, one which is efficient under $\mathbb{H}_0$ but inconsistent under $\mathbb{H}_1$, and another which is consistent under $\mathbb{H}_1$, then construct a test as a quadratic form in the differences of the estimators. In the case of testing a hypothesis $\mathbb{H}_0 : r(\beta) = \theta_0$ let $\widehat{\beta}_{\text{ols}}$ denote the unconstrained least squares estimator and let $\widetilde{\beta}_{\text{emd}}$ denote the efficient minimum distance estimator which imposes $r(\beta) = \theta_0$. Both estimators are consistent under $\mathbb{H}_0$ but $\widetilde{\beta}_{\text{emd}}$ is asymptotically efficient. Under $\mathbb{H}_1$, $\widehat{\beta}_{\text{ols}}$ is consistent for β but $\widetilde{\beta}_{\text{emd}}$ is inconsistent. The difference has the asymptotic distribution

$$\sqrt{n}\left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right) \xrightarrow[d]{} \mathrm{N}\left(0, \boldsymbol{V}_\beta \boldsymbol{R}\left(\boldsymbol{R}'\boldsymbol{V}_\beta \boldsymbol{R}\right)^{-1} \boldsymbol{R}'\boldsymbol{V}_\beta\right).$$

Let $\boldsymbol{A}^-$ denote the Moore-Penrose generalized inverse. The Hausman statistic for $\mathbb{H}_0$ is

$$\begin{aligned} H &= \left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right)' \widehat{\text{avar}}\left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right)^- \left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right) \\ &= n\left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right)' \left(\widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta \widehat{\boldsymbol{R}}\right)^{-1} \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\right)^- \left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right). \end{aligned}$$

[1] All coefficients except the intercept.

The matrix $\widehat{\boldsymbol{V}}_\beta^{1/2}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta^{1/2}$ idempotent so its generalized inverse is itself. (See Section A.11.) It follows that

$$\begin{aligned}\left(\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\right)^{-} &= \widehat{\boldsymbol{V}}_\beta^{-1/2}\left(\widehat{\boldsymbol{V}}_\beta^{1/2}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta^{1/2}\right)^{-}\widehat{\boldsymbol{V}}_\beta^{-1/2} \\ &= \widehat{\boldsymbol{V}}_\beta^{-1/2}\widehat{\boldsymbol{V}}_\beta^{1/2}\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta^{1/2}\widehat{\boldsymbol{V}}_\beta^{-1/2} \\ &= \widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'.\end{aligned}$$

Thus the Hausman statistic is

$$H = n\left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right)'\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right).$$

In the context of linear restrictions, $\widehat{\boldsymbol{R}} = \boldsymbol{R}$ and $\boldsymbol{R}'\widetilde{\beta} = \theta_0$ so the statistic takes the form

$$H = n\left(\boldsymbol{R}'\widehat{\beta}_{\text{ols}} - \theta_0\right)'\widehat{\boldsymbol{R}}\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_\beta\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\beta}_{\text{ols}} - \theta_0\right),$$

which is precisely the Wald statistic. With nonlinear restrictions W and H can differ.

In either case we see that that the asymptotic null distribution of the Hausman statistic H is χ^2_q, so the appropriate test is to reject $\mathbb{H}_0$ in favor of $\mathbb{H}_1$ if $H > c$ where c is a critical value taken from the χ^2_q distribution.

Theorem 9.7 For general hypotheses the Hausman test statistic is

$$H = n\left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right)'\widehat{\boldsymbol{R}}\left(\widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_\beta\widehat{\boldsymbol{R}}\right)^{-1}\widehat{\boldsymbol{R}}'\left(\widehat{\beta}_{\text{ols}} - \widetilde{\beta}_{\text{emd}}\right).$$

Under Assumptions 7.2, 7.3, 7.4, and $\mathbb{H}_0 : r(\beta) = \theta_0 \in \mathbb{R}^q$, $H \xrightarrow[d]{} \chi^2_q$.

9.16 Score Tests

Score tests are traditionally derived in likelihood analysis but can more generally be constructed from first-order conditions evaluated at restricted estimates. We focus on the likelihood derivation.

Given the log likelihood function $\ell_n(\beta, \sigma^2)$, a restriction $\mathbb{H}_0 : r(\beta) = \theta_0$, and restricted estimators $\widetilde{\beta}$ and $\widetilde{\sigma}^2$, the **score statistic** for $\mathbb{H}_0$ is defined as

$$S = \left(\frac{\partial}{\partial\beta}\ell_n(\widetilde{\beta}, \widetilde{\sigma}^2)\right)'\left(-\frac{\partial^2}{\partial\beta\partial\beta'}\ell_n(\widetilde{\beta}, \widetilde{\sigma}^2)\right)^{-1}\left(\frac{\partial}{\partial\beta}\ell_n(\widetilde{\beta}, \widetilde{\sigma}^2)\right).$$

The idea is that if the restriction is true then the restricted estimators should be close to the maximum of the log-likelihood where the derivative is zero. However if the restriction is false then the restricted estimators should be distant from the maximum and the derivative should be large. Hence small values of S are expected under $\mathbb{H}_0$ and large values under $\mathbb{H}_1$. Tests of $\mathbb{H}_0$ reject for large values of S.

We explore the score statistic in the context of the normal regression model and linear hypotheses $r(\beta) = \boldsymbol{R}'\beta$. Recall that in the normal regression log-likelihood function is

$$\ell_n(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2.$$

The constrained MLE under linear hypotheses is constrained least squares

$$\begin{aligned}\widetilde{\beta} &= \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\left[\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c}\right)\\ \widetilde{e}_i &= Y_i - X_i'\widetilde{\beta}\\ \widetilde{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\widetilde{e}_i^2.\end{aligned}$$

We can calculate that the derivative and Hessian are

$$\begin{aligned}\frac{\partial}{\partial\beta}\ell_n(\widetilde{\beta},\widetilde{\sigma}^2) &= \frac{1}{\widetilde{\sigma}^2}\sum_{i=1}^{n}X_i\left(Y_i - X_i'\widetilde{\beta}\right) = \frac{1}{\widetilde{\sigma}^2}\boldsymbol{X}'\widetilde{\boldsymbol{e}}\\ -\frac{\partial^2}{\partial\beta\partial\beta'}\ell_n(\widetilde{\beta},\widetilde{\sigma}^2) &= \frac{1}{\widetilde{\sigma}^2}\sum_{i=1}^{n}X_iX_i' = \frac{1}{\widetilde{\sigma}^2}\boldsymbol{X}'\boldsymbol{X}.\end{aligned}$$

Since $\widetilde{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\widetilde{\beta}$ we can further calculate that

$$\begin{aligned}\frac{\partial}{\partial\beta}\ell_n(\widetilde{\beta},\widetilde{\sigma}^2) &= \frac{1}{\widetilde{\sigma}^2}\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\widetilde{\beta}\right)\\ &= \frac{1}{\widetilde{\sigma}^2}\left(\boldsymbol{X}'\boldsymbol{X}\right)\left(\widehat{\beta} - \widetilde{\beta}\right)\\ &= \frac{1}{\widetilde{\sigma}^2}\boldsymbol{R}\left[\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right]^{-1}\left(\boldsymbol{R}'\widehat{\beta} - \boldsymbol{c}\right).\end{aligned}$$

Together we find that

$$S = \left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)'\left(\boldsymbol{R}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{R}\right)^{-1}\left(\boldsymbol{R}'\widehat{\boldsymbol{\beta}} - \boldsymbol{c}\right)/\widetilde{\sigma}^2.$$

This is identical to the homoskedastic Wald statistic with s^2 replaced by $\widetilde{\sigma}^2$. We can also write S as a monotonic transformation of the F statistic, as

$$S = n\frac{\left(\widetilde{\sigma}^2 - \widehat{\sigma}^2\right)}{\widetilde{\sigma}^2} = n\left(1 - \frac{\widehat{\sigma}^2}{\widetilde{\sigma}^2}\right) = n\left(1 - \frac{1}{1 + \frac{q}{n-k}F}\right).$$

The test "Reject $\mathbb{H}_0$ for large values of S" is identical to the test "Reject $\mathbb{H}_0$ for large values of F" so they are identical tests. Since for the normal regression model the exact distribution of F is known, it is better to use the F statistic with F p-values.

In more complicated settings a potential advantage of score tests is that they are calculated using the restricted parameter estimates $\widetilde{\beta}$ rather than the unrestricted estimates $\widehat{\beta}$. Thus when $\widetilde{\beta}$ is relatively easy to calculate there can be a preference for score statistics. This is not a concern for linear restrictions.

More generally, score and score-like statistics can be constructed from first-order conditions evaluated at restricted parameter estimates. Also, when test statistics are constructed using covariance matrix estimators which are calculated using restricted parameter estimates (e.g. restricted residuals) then these are often described as score tests.

An example of the latter is the Wald-type statistic

$$W = \left(r(\widehat{\beta}) - \theta_0\right)'\left(\widehat{\boldsymbol{R}}'\widetilde{\boldsymbol{V}}_{\widehat{\beta}}\widehat{\boldsymbol{R}}\right)^{-1}\left(r(\widehat{\beta}) - \theta_0\right)$$

where the covariance matrix estimate $\widetilde{\boldsymbol{V}}_{\widehat{\beta}}$ is calculated using the restricted residuals $\widetilde{e}_i = Y_i - X_i'\widetilde{\beta}$. This may be a good choice when β and θ are high-dimensional as in this context there may be worry that the estimator $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ is imprecise.

9.17 Problems with Tests of Nonlinear Hypotheses

While the t and Wald tests work well when the hypothesis is a linear restriction on β, 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 $Y \sim \mathrm{N}(\beta, \sigma^2)$ and consider the hypothesis $\mathbb{H}_0 : \beta = 1$. Let $\widehat{\beta}$ and $\widehat{\sigma}^2$ be the sample mean and variance of Y. The standard Wald statistic to test $\mathbb{H}_0$ is

$$W = n\frac{\left(\widehat{\beta} - 1\right)^2}{\widehat{\sigma}^2}.$$

Notice that $\mathbb{H}_0$ is equivalent to the hypothesis $\mathbb{H}_0(s) : \beta^s = 1$ for any positive integer s. Letting $r(\beta) = \beta^s$, and noting $\boldsymbol{R} = s\beta^{s-1}$, we find that the Wald statistic to test $\mathbb{H}_0(s)$ is

$$W_s = n\frac{\left(\widehat{\beta}^s - 1\right)^2}{\widehat{\sigma}^2 s^2 \widehat{\beta}^{2s-2}}.$$

While the hypothesis $\beta^s = 1$ is unaffected by the choice of s, the statistic W_s varies with s. This is an unfortunate feature of the Wald statistic.

To demonstrate this effect, we have plotted in Figure 9.2 the Wald statistic W_s as a function of s, setting $n/\widehat{\sigma}^2 = 10$. The increasing line is for the case $\widehat{\beta} = 0.8$. The decreasing line is for the case $\widehat{\beta} = 1.6$. It is easy to see that in each case there are values of s for which the test statistic is significant relative to asymptotic critical values, while there are other values of s for which the test statistic is insignificant. This is distressing because the choice of s is arbitrary and irrelevant to the actual hypothesis.

Our first-order asymptotic theory is not useful to help pick s, as $W_s \underset{d}{\longrightarrow} \chi^2_1$ under $\mathbb{H}_0$ for any s. This is a context where **Monte Carlo simulation** can be quite useful as a tool to study and compare the exact distributions of statistical procedures in finite samples. The method uses random simulation to create artificial datasets to which we apply the statistical tools of interest. This produces random draws from the statistic's sampling distribution. Through repetition, features of this distribution can be calculated.

In the present context of the Wald statistic, one feature of importance is the Type I error of the test using the asymptotic 5% critical value 3.84 – the probability of a false rejection, $\mathbb{P}\left[W_s > 3.84 \mid \beta = 1\right]$. Given the simplicity of the model this probability depends only on s, n, and σ^2. In Table 9.2 we report the results of a Monte Carlo simulation where we vary these three parameters. The value of s is varied from 1 to 10, n is varied among 20, 100, and 500, and σ is varied among 1 and 3. The table reports the simulation estimate of the Type I error probability from 50,000 random samples. Each row of the table corresponds to a different value of s – and thus corresponds to a particular choice of test statistic. The second through seventh columns contain the Type I error probabilities for different combinations of n and σ. These probabilities are calculated as the percentage of the 50,000 simulated Wald statistics W_s which are larger than 3.84. The null hypothesis $\beta^s = 1$ is true so these probabilities are Type I error.

To interpret the table remember that the ideal Type I error probability is 5% (.05) with deviations indicating distortion. Type I error rates between 3% and 8% are considered reasonable. Error rates above 10% are considered excessive. Rates above 20% are unacceptable. When comparing statistical procedures we compare the rates row by row, looking for tests for which rejection rates are close to 5% and rarely fall outside of the 3%-8% range. For this particular example the only test which meets this criterion is the conventional $W = W_1$ test. Any other s leads to a test with unacceptable Type I error probabilities.

In Table 9.2 you can also see the impact of variation in sample size. In each case the Type I error probability improves towards 5% as the sample size n increases. There is, however, no magic choice of n for which all tests perform uniformly well. Test performance deteriorates as s increases which is not surprising given the dependence of W_s on s as shown in Figure 9.2.

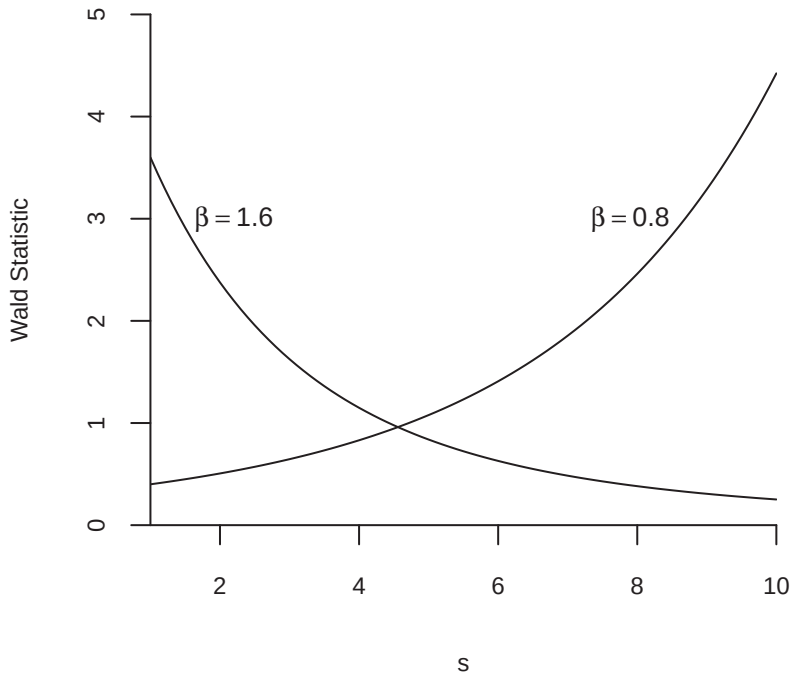


Figure 9.2: Wald Statistic as a Function of s

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 obvious and the best choices may be counter-intuitive.

This point can be illustrated through an example based on Gregory and Veall (1985). Take the model

$$\begin{aligned} Y &= \beta_0 + X_1\beta_1 + X_2\beta_2 + e \\ \mathbb{E}[Xe] &= 0 \end{aligned} \tag{9.14}$$

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 $\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ be the least squares estimator of (9.14), let $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ be an estimator of the covariance matrix for $\widehat{\beta}$ and set $\widehat{\theta} = \widehat{\beta}_1/\widehat{\beta}_2$. Define

$$\widehat{\boldsymbol{R}}_1 = \begin{pmatrix} 0 \\ \dfrac{1}{\widehat{\beta}_2} \\ -\dfrac{\widehat{\beta}_1}{\widehat{\beta}_2^2} \end{pmatrix}$$

Table 9.2: Type I Error Probability of Asymptotic 5% $W(s)$ Test

s	$\sigma = 1$			$\sigma = 3$		
	$n = 20$	$n = 100$	$n = 500$	$n = 20$	$n = 100$	$n = 500$
1	0.05	0.05	0.05	0.05	0.05	0.05
2	0.07	0.06	0.05	0.14	0.08	0.06
3	0.09	0.06	0.05	0.21	0.12	0.07
4	0.12	0.07	0.05	0.25	0.15	0.08
5	0.14	0.08	0.06	0.27	0.18	0.10
6	0.16	0.09	0.06	0.30	0.20	0.12
7	0.18	0.10	0.06	0.32	0.22	0.13
8	0.20	0.12	0.07	0.33	0.24	0.14
9	0.21	0.13	0.07	0.34	0.25	0.16
10	0.23	0.14	0.08	0.35	0.26	0.17

Rejection frequencies from 50,000 simulated random samples.

so that the standard error for $\widehat{\theta}$ is $s(\widehat{\theta}) = \left(\widehat{\boldsymbol{R}}_1' \widehat{\boldsymbol{V}}_{\widehat{\beta}} \widehat{\boldsymbol{R}}_1\right)^{1/2}$. In this case a t-statistic for $\mathbb{H}_0$ is

$$T_1 = \frac{\left(\frac{\widehat{\beta}_1}{\widehat{\beta}_2} - \theta_0\right)}{s(\widehat{\theta})}.$$

An alternative statistic can be constructed through reformulating the null hypothesis as

$$\mathbb{H}_0 : \beta_1 - \theta_0 \beta_2 = 0.$$

A t-statistic based on this formulation of the hypothesis is

$$T_2 = \frac{\widehat{\beta}_1 - \theta_0 \widehat{\beta}_2}{\left(\boldsymbol{R}_2' \widehat{\boldsymbol{V}}_{\widehat{\beta}} \boldsymbol{R}_2\right)^{1/2}}$$

where

$$\boldsymbol{R}_2 = \begin{pmatrix} 0 \\ 1 \\ -\theta_0 \end{pmatrix}.$$

To compare T_1 and T_2 we perform another simple Monte Carlo simulation. We let X_1 and X_2 be mutually independent N(0, 1) variables, e be an independent 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 0.1, 0.25, 0.50, 0.75, and 1.0 and n among 100 and 500.

The one-sided Type I error probabilities $\mathbb{P}[T < -1.645]$ and $\mathbb{P}[T > 1.645]$ are calculated from 50,000 simulated samples. The results are presented in Table 9.3. Ideally, the entries in the table should be 0.05. However, the rejection rates for the T_1 statistic diverge greatly from this value, especially for small values of β_2. The left tail probabilities $\mathbb{P}[T_1 < -1.645]$ greatly exceed 5%, while the right tail probabilities $\mathbb{P}[T_1 > 1.645]$ are close to zero in most cases. In contrast, the rejection rates for the T_2 statistic are invariant to the value of β_2 and equal 5% for both sample sizes. The implication of Table 9.3 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.

Table 9.3: Type I Error Probability of Asymptotic 5% t-tests

β_2	$n = 100$				$n = 500$			
	$\mathbb{P}(T < -1.645)$		$\mathbb{P}(T > 1.645)$		$\mathbb{P}(T < -1.645)$		$\mathbb{P}(T > 1.645)$	
	T_1	T_2	T_1	T_2	T_1	T_2	T_1	T_2
0.10	0.47	0.05	0.00	0.05	0.28	0.05	0.00	0.05
0.25	0.27	0.05	0.00	0.05	0.16	0.05	0.00	0.05
0.50	0.14	0.05	0.00	0.05	0.12	0.05	0.00	0.05
0.75	0.03	0.05	0.00	0.05	0.08	0.05	0.01	0.05
1.00	0.00	0.05	0.00	0.05	0.03	0.05	0.03	0.05

Rejection frequencies from 50,000 simulated random samples.

A simple solution is to use the minimum distance statistic J which equals W with $r = 1$ in the first example, and $|T_2|$ in the second example. The minimum distance statistic is invariant to the algebraic formulation of the null hypothesis so is immune to this problem. Whenever possible, the Wald statistic should not be used to test nonlinear hypotheses.

Theoretical investigations of these issues include Park and Phillips (1988) and Dufour (1997).

9.18 Monte Carlo Simulation

In Section 9.17 we introduced the method of Monte Carlo simulation to illustrate the small sample problems with tests of nonlinear hypotheses. In this section we describe the method in more detail.

Recall, our data consist of observations (Y_i, X_i) which are random draws from a population distribution F. Let θ be a parameter and let $T = T((Y_1, X_1), ..., (Y_n, X_n), \theta)$ be a statistic of interest, for example an estimator $\widehat{\theta}$ or a t-statistic $(\widehat{\theta} - \theta)/s(\widehat{\theta})$. The exact distribution of T is

$$G(u, F) = \mathbb{P}[T \leq u \mid F].$$

While the asymptotic distribution of T might be known, the exact (finite sample) distribution G is generally unknown.

Monte Carlo simulation uses numerical simulation to compute $G(u, F)$ for selected choices of F. This is useful to investigate the performance of the statistic T in reasonable situations and sample sizes. The basic idea is that for any given F the distribution function $G(u, F)$ can be calculated numerically through simulation. The name **Monte Carlo** derives from the Mediterranean gambling resort where games of chance are played.

The method of Monte Carlo is simple to describe. The researcher chooses F (the distribution of the pseudo data) and the sample size n. A "true" value of θ is implied by this choice, or equivalently the value θ 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 $\left(Y_i^*, X_i^*\right)$, $i = 1, ..., n$, are drawn from the distribution F using the computer's random number generator.
2. The statistic $T = T\left(\left(Y_1^*, X_1^*\right), ..., \left(Y_n^*, X_n^*\right), \theta\right)$ is calculated on this pseudo data.

For step 1, computer packages have built-in random number procedures including U[0, 1] and N(0, 1). From these most random variables can be constructed. (For example, a chi-square can be generated by sums of squares of normals.)

For step 2, it is important that the statistic be evaluated at the "true" value of θ corresponding to the choice of F.

The above experiment creates one random draw T from the distribution $G(u,F)$. This is one observation from an unknown distribution. Clearly, from one observation very little can be said. So the researcher repeats the experiment B times where B is a large number. Typically, we set $B \geq 1000$. We will discuss this choice later.

Notationally, let the b^{th} experiment result in the draw T_b, $b = 1,...,B$. These results are stored. After all B experiments have been calculated these results constitute a random sample of size B from the distribution of $G(u,F) = \mathbb{P}[T_b \leq u] = \mathbb{P}[T \leq u \mid F]$.

From a random sample we can estimate any feature of interest using (typically) a method of moments estimator. We now describe some specific examples.

Suppose we are interested in the bias, mean-squared error (MSE), and/or variance of the distribution of $\widehat{\theta} - \theta$. We then set $T = \widehat{\theta} - \theta$, run the above experiment, and calculate

$$\widehat{\text{bias}}\left[\widehat{\theta}\right] = \frac{1}{B}\sum_{b=1}^{B} T_b = \frac{1}{B}\sum_{b=1}^{B} \widehat{\theta}_b - \theta$$

$$\widehat{\text{mse}}\left[\widehat{\theta}\right] = \frac{1}{B}\sum_{b=1}^{B} (T_b)^2 = \frac{1}{B}\sum_{b=1}^{B} \left(\widehat{\theta}_b - \theta\right)^2$$

$$\widehat{\text{var}}\left[\widehat{\theta}\right] = \widehat{\text{mse}}\left[\widehat{\theta}\right] - \left(\widehat{\text{bias}}\left[\hat{\theta}\right]\right)^2$$

Suppose we are interested in the Type I error associated with an asymptotic 5% two-sided t-test. We would then set $T = \left|\widehat{\theta} - \theta\right| / s(\widehat{\theta})$ and calculate

$$\widehat{P} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{T_b \geq 1.96\}, \tag{9.15}$$

the percentage of the simulated t-ratios which exceed the asymptotic 5% critical value.

Suppose we are interested in the 5% and 95% quantile of $T = \widehat{\theta}$ or $T = \left(\widehat{\theta} - \theta\right) / s(\widehat{\theta})$. We then compute the 5% and 95% sample quantiles of the sample $\{T_b\}$. For details on quantile estimation see Section 11.13 of *Probability and Statistics for Economists*.

The typical purpose of a Monte Carlo simulation is to investigate the performance of a statistical procedure in realistic settings. Generally, the performance will depend on n and F. In many cases an estimator or test may perform wonderfully for some values and poorly for others. It is therefore useful to conduct a variety of experiments for a selection of choices of n and F.

As discussed above the researcher must select the number of experiments B. Often this is called the number of **replications**. Quite simply, a larger B results in more precise estimates of the features of interest of G but requires more computational time. In practice, therefore, the choice of B is often guided by the computational demands of the statistical procedure. Since the results of a Monte Carlo experiment are estimates computed from a random sample of size B it is straightforward to calculate standard errors for any quantity of interest. If the standard error is too large to make a reliable inference then B will have to be increased. A useful rule-of-thumb is to set $B = 10,000$ whenever possible.

In particular, it is simple to make inferences about rejection probabilities from statistical tests, such as the percentage estimate reported in (9.15). The random variable $\mathbb{1}\{T_b \geq 1.96\}$ is i.i.d. Bernoulli, equalling 1 with probability $p = \mathbb{E}[\mathbb{1}\{T_b \geq 1.96\}]$. The average (9.15) is therefore an unbiased estimator of p with standard error $s\left(\widehat{p}\right) = \sqrt{p\left(1-p\right)/B}$. As p is unknown, this may be approximated by replacing p with $\widehat{p}$ or with an hypothesized value. For example, if we are assessing an asymptotic 5% test, then we can set $s\left(\widehat{p}\right) = \sqrt{(.05)(.95)/B} \simeq .22/\sqrt{B}$. Hence, standard errors for $B = 100$, 1000, and 5000, are, respectively, $s\left(\widehat{p}\right) = .022$, .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 may not be exactly possible in the case of simulation results as they are inherently random. For example suppose a paper investigates a statistical test and reports a simulated rejection probability of 0.07 based on a simulation with $B = 100$ replications. Suppose you attempt to replicate this result and find a rejection probability of 0.03 (again using $B = 100$ simulation replications). Should you conclude that you have failed in your attempt? Absolutely not! Under the hypothesis that both simulations are identical you have two independent estimates, $\widehat{p}_1 = 0.07$ and $\widehat{p}_2 = 0.03$, of a common probability p. The asymptotic (as $B \to \infty$) distribution of their difference is $\sqrt{B}\left(\widehat{p}_1 - \widehat{p}_2\right) \underset{d}{\longrightarrow} \mathrm{N}(0, 2p(1-p))$, so a standard error for $\widehat{p}_1 - \widehat{p}_2 = 0.04$ is $\widehat{s} = \sqrt{2\overline{p}(1-\overline{p})/B} \simeq 0.03$, using the estimate $\overline{p} = (\widehat{p}_1 + \widehat{p}_2)/2$. Since the t-ratio $0.04/0.03 = 1.3$ is not statistically significant it is incorrect to reject the null hypothesis that the two simulations are identical. The difference between the results $\widehat{p}_1 = 0.07$ and $\widehat{p}_2 = 0.03$ is consistent with random variation.

What should be done? The first mistake was to copy the previous paper's choice of $B = 100$. Instead, suppose you set $B = 10,000$ and now obtain $\widehat{p}_2 = 0.04$. Then $\widehat{p}_1 - \widehat{p}_2 = 0.03$ and a standard error is $\widehat{s} = \sqrt{\overline{p}(1-\overline{p})\,(1/100 + 1/10000)} \simeq 0.02$. 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. Most do not have similar policies regarding simulations. Nevertheless, it is good professional practice to make your simulations available. The best practice is to post your simulation code on your webpage. This invites others to build on and use your results, leading to possible collaboration, citation, and/or advancement.

9.19 Confidence Intervals by Test Inversion

There is a close relationship between hypothesis tests and confidence intervals. We observed in Section 7.13 that the standard 95% asymptotic confidence interval for a parameter θ is

$$\widehat{C} = \left[\widehat{\theta} - 1.96 \times s(\widehat{\theta}),\quad \widehat{\theta} + 1.96 \times s(\widehat{\theta})\right] = \{\theta : |T(\theta)| \le 1.96\}. \tag{9.16}$$

That is, we can describe $\widehat{C}$ as "The point estimate plus or minus 2 standard errors" or "The set of parameter values not rejected by a two-sided t-test." The second definition, known as **test statistic inversion**, is a general method for finding confidence intervals, and typically produces confidence intervals with excellent properties.

Given a test statistic $T(\theta)$ and critical value c, the acceptance region "Accept if $T(\theta) \le c$" is identical to the confidence interval $\widehat{C} = \{\theta : T(\theta) \le c\}$. Since the regions are identical the probability of coverage $\mathbb{P}\left[\theta \in \widehat{C}\right]$ equals the probability of correct acceptance $\mathbb{P}\left[\text{Accept} \mid \theta\right]$ 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 $\theta = r(\beta)$ is a nonlinear function of the coefficient vector β. In this case the standard confidence interval for θ is the set $\widehat{C}$ as in (9.16) where $\widehat{\theta} = r(\widehat{\beta})$ is the point estimator and $s(\widehat{\theta}) = \sqrt{\widehat{\boldsymbol{R}}' \widehat{\boldsymbol{V}}_{\widehat{\beta}} \widehat{\boldsymbol{R}}}$ is the delta method standard error. This confidence interval is inverting the t-test based on the nonlinear hypothesis $r(\beta) = \theta$. The trouble is that in Section 9.17 we learned that there is no unique t-statistic for tests of nonlinear hypotheses and that the choice of parameterization matters greatly.

For example, if $\theta = \beta_1/\beta_2$ then the coverage probability of the standard interval (9.16) is 1 minus the probability of the Type I error, which as shown in Table 8.2 can be far from the nominal 5%.

In this example a good solution is the same as discussed in Section 9.17 – to rewrite the hypothesis as a linear restriction. The hypothesis $\theta = \beta_1/\beta_2$ is the same as $\theta\beta_2 = \beta_1$. The t-statistic for this restriction is

$$T(\theta) = \frac{\widehat{\beta}_1 - \widehat{\beta}_2\theta}{\left(\boldsymbol{R}'\widehat{\boldsymbol{V}}_{\widehat{\beta}}\boldsymbol{R}\right)^{1/2}}$$

where

$$\boldsymbol{R} = \begin{pmatrix} 1 \\ -\theta \end{pmatrix}$$

and $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ is the covariance matrix for $(\widehat{\beta}_1\ \widehat{\beta}_2)$. A 95% confidence interval for $\theta = \beta_1/\beta_2$ is the set of values of θ such that $|T(\theta)| \le 1.96$. Since $T(\theta)$ is a nonlinear function of θ one method to find the confidence set is grid search over θ.

For example, in the wage equation

$$\log(\mathit{wage}) = \beta_1 \mathit{experience} + \beta_2 \mathit{experience}^2/100 + \cdots$$

the highest expected wage occurs at $\mathit{experience} = -50\beta_1/\beta_2$. From Table 4.1 we have the point estimate $\widehat{\theta} = 29.8$ and we can calculate the standard error $s(\widehat{\theta}) = 0.022$ for a 95% confidence interval [29.8, 29.9]. However, if we instead invert the linear form of the test we numerically find the interval [29.1, 30.6] which is much larger. From the evidence presented in Section 9.17 we know the first interval can be quite inaccurate and the second interval is greatly preferred.

9.20 Multiple Tests and Bonferroni Corrections

In most applications economists examine a large number of estimates, test statistics, and p-values. What does it mean (or does it mean anything) if one statistic appears to be "significant" after examining a large number of statistics? This is known as the problem of **multiple testing** or **multiple comparisons**.

To be specific, suppose we examine a set of k coefficients, standard errors and t-ratios, and consider the "significance" of each statistic. Based on conventional reasoning, for each coefficient we would reject the hypothesis that the coefficient is zero with asymptotic size α if the absolute t-statistic exceeds the $1-\alpha$ critical value of the normal distribution, or equivalently if the p-value for the t-statistic is smaller than α. If we observe that one of the k statistics is "significant" based on this criterion, that means that one of the p-values is smaller than α, or equivalently, that the smallest p-value is smaller than α. We can then rephrase the question: Under the joint hypothesis that a set of k hypotheses are all true, what is the probability that the smallest p-value is smaller than α? In general, we cannot provide a precise answer to this quesion, but the Bonferroni correction bounds this probability by αk. The Bonferroni method furthermore suggests that if we want the **familywise error probability** (the probability that one of the tests falsely rejects) to be bounded below α, then an appropriate rule is to reject only if the smallest p-value is smaller than α/k. Equivalently, the Bonferroni familywise p-value is $k\min_{j\le k} p_j$.

Formally, suppose we have k hypotheses $\mathbb{H}_j$, $j = 1, ..., k$. For each we have a test and associated p-value p_j with the property that when $\mathbb{H}_j$ is true $\lim_{n\to\infty}\mathbb{P}\left[p_j < \alpha\right] = \alpha$. We then observe that among the k tests, one of the k is "significant" if $\min_{j\le k} p_j < \alpha$. This event can be written as

$$\left\{\min_{j\le k} p_j < \alpha\right\} = \bigcup_{j=1}^{k}\left\{p_j < \alpha\right\}.$$

Boole's inequality states that for any k events A_j, $\mathbb{P}\left[\bigcup_{j=1}^{k} A_j\right] \leq \sum_{j=1}^{k} \mathbb{P}[A_k]$. Thus

$$\mathbb{P}\left[\min_{j\leq k} p_j < \alpha\right] \leq \sum_{j=1}^{k} \mathbb{P}\left[p_j < \alpha\right] \to k\alpha$$

as stated. This demonstates that the asymptotic familywise rejection probability is at most k times the individual rejection probability.

Furthermore,

$$\mathbb{P}\left[\min_{j\leq k} p_j < \frac{\alpha}{k}\right] \leq \sum_{j=1}^{k} \mathbb{P}\left[p_j < \frac{\alpha}{k}\right] \to \alpha.$$

This demonstrates that the asymptotic familywise rejection probability can be controlled (bounded below α) if each individual test is subjected to the stricter standard that a p-value must be smaller than α/k to be labeled as "significant".

To illustrate, suppose we have two coefficient estimates with individual p-values 0.04 and 0.15. Based on a conventional 5% level the standard individual tests would suggest that the first coefficient estimate is "significant" but not the second. A Bonferroni 5% test, however, does not reject as it would require that the smallest p-value be smaller than 0.025, which is not the case in this example. Alternatively, the Bonferroni familywise p-value is $0.04 \times 2 = 0.08$, which is not significant at the 5% level.

In contrast, if the two p-values were 0.01 and 0.15, then the Bonferroni familywise p-value would be $0.01 \times 2 = 0.02$, which is significant at the 5% level.

9.21 Power and Test Consistency

The **power** of a test is the probability of rejecting $\mathbb{H}_0$ when $\mathbb{H}_1$ is true.

For simplicity suppose that Y_i is i.i.d. $\mathrm{N}(\theta, \sigma^2)$ with σ^2 known, consider the t-statistic $T(\theta) = \sqrt{n}\left(\overline{Y} - \theta\right)/\sigma$, and tests of $\mathbb{H}_0 : \theta = 0$ against $\mathbb{H}_1 : \theta > 0$. We reject $\mathbb{H}_0$ if $T = T(0) > c$. Note that

$$T = T(\theta) + \sqrt{n}\theta/\sigma$$

and $T(\theta)$ has an exact $\mathrm{N}(0,1)$ distribution. This is because $T(\theta)$ is centered at the true mean θ, while the test statistic $T(0)$ is centered at the (false) hypothesized mean of 0.

The power of the test is

$$\mathbb{P}\left[T > c \mid \theta\right] = \mathbb{P}\left[\mathrm{Z} + \sqrt{n}\theta/\sigma > c\right] = 1 - \Phi\left(c - \sqrt{n}\theta/\sigma\right).$$

This function is monotonically increasing in μ and n, and decreasing in σ and c.

Notice that for any c and $\theta \neq 0$ the power increases to 1 as $n \to \infty$. This means that for $\theta \in \mathbb{H}_1$ the test will reject $\mathbb{H}_0$ with probability approaching 1 as the sample size gets large. We call this property **test consistency**.

Definition 9.3 A test of $\mathbb{H}_0 : \theta \in \Theta_0$ is **consistent against fixed alternatives** if for all $\theta \in \Theta_1$, $\mathbb{P}\left[\text{Reject } \mathbb{H}_0 \mid \theta\right] \to 1$ as $n \to \infty$.

For tests of the form "Reject $\mathbb{H}_0$ if $T > c$", a sufficient condition for test consistency is that the T diverges to positive infinity with probability one for all $\theta \in \Theta_1$.

Definition 9.4 We say that $T \xrightarrow[p]{} \infty$ as $n \to \infty$ if for all $M < \infty$, $\mathbb{P}[T \leq M] \to 0$ as $n \to \infty$. Similarly, we say that $T \xrightarrow[p]{} -\infty$ as $n \to \infty$ if for all $M < \infty$, $\mathbb{P}[T \geq -M] \to 0$ as $n \to \infty$.

In general, t-tests and Wald tests are consistent against fixed alternatives. Take a t-statistic for a test of $\mathbb{H}_0 : \theta = \theta_0$, $T = \left(\widehat{\theta} - \theta_0\right) / s(\widehat{\theta})$ where θ_0 is a known value and $s(\widehat{\theta}) = \sqrt{n^{-1}\widehat{V}_\theta}$. Note that

$$T = \frac{\widehat{\theta} - \theta}{s(\widehat{\theta})} + \frac{\sqrt{n}(\theta - \theta_0)}{\sqrt{\widehat{V}_\theta}}.$$

The first term on the right-hand-side converges in distribution to $\mathrm{N}(0,1)$. The second term on the right-hand-side equals zero if $\theta = \theta_0$, converges in probability to $+\infty$ if $\theta > \theta_0$, and converges in probability to $-\infty$ if $\theta < \theta_0$. Thus the two-sided t-test is consistent against $\mathbb{H}_1 : \theta \neq \theta_0$, and one-sided t-tests are consistent against the alternatives for which they are designed.

Theorem 9.8 Under Assumptions 7.2, 7.3, and 7.4, for $\theta = r(\beta) \neq \theta_0$ and $q = 1$, then $|T| \xrightarrow[p]{} \infty$. For any $c < \infty$ the test "Reject $\mathbb{H}_0$ if $|T| > c$" is consistent against fixed alternatives.

The Wald statistic for $\mathbb{H}_0 : \theta = r(\beta) = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ is $W = n\left(\widehat{\theta} - \theta_0\right)' \widehat{\boldsymbol{V}}_\theta^{-1} \left(\widehat{\theta} - \theta_0\right)$. Under $\mathbb{H}_1$, $\widehat{\theta} \xrightarrow[p]{} \theta \neq \theta_0$. Thus $\left(\widehat{\theta} - \theta_0\right)' \widehat{\boldsymbol{V}}_\theta^{-1} \left(\widehat{\theta} - \theta_0\right) \xrightarrow[p]{} (\theta - \theta_0)' \boldsymbol{V}_\theta^{-1} (\theta - \theta_0) > 0$. Hence under $\mathbb{H}_1$, $W \xrightarrow[p]{} \infty$. Again, this implies that Wald tests are consistent.

Theorem 9.9 Under Assumptions 7.2, 7.3, and 7.4, for $\theta = r(\beta) \neq \theta_0$, then $W \xrightarrow[p]{} \infty$. For any $c < \infty$ the test "Reject $\mathbb{H}_0$ if $W > c$" is consistent against fixed alternatives.

9.22 Asymptotic Local Power

Consistency is a good property for a test but is does not provided a tool to calculate test power. To approximate the power function we need a distributional approximation.

The standard asymptotic method for power analysis uses what are called **local alternatives**. This is similar to our analysis of restriction estimation under misspecification (Section 8.13). The technique is to index the parameter by sample size so that the asymptotic distribution of the statistic is continuous in a localizing parameter. In this section we consider t-tests on real-valued parameters and in the next section Wald tests. Specifically, we consider parameter vectors β_n which are indexed by sample size n and satisfy the real-valued relationship

$$\theta_n = r(\beta_n) = \theta_0 + n^{-1/2} h \tag{9.17}$$

where the scalar h is called a **localizing parameter**. We index β_n and θ_n by sample size to indicate their dependence on n. The way to think of (9.17) is that the true value of the parameters are β_n and θ_n. The parameter θ_n is close to the hypothesized value θ_0, with deviation $n^{-1/2}h$.

The specification (9.17) states that for any fixed h, θ_n approaches θ_0 as n gets large. Thus θ_n is "close" or "local" to θ_0. The concept of a localizing sequence (9.17) might seem odd since in the actual world the sample size cannot mechanically affect the value of the parameter. Thus (9.17) should not be interpreted literally. Instead, it should be interpreted as a technical device which allows the asymptotic distribution to be continuous in the alternative hypothesis.

To evaluate the asymptotic distribution of the test statistic we start by examining the scaled estimator centered at the hypothesized value θ_0. Breaking it into a term centered at the true value θ_n and a remainder we find

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right)=\sqrt{n}\left(\widehat{\theta}-\theta_n\right)+\sqrt{n}\left(\theta_n-\theta_0\right)=\sqrt{n}\left(\widehat{\theta}-\theta_n\right)+h$$

where the second equality is (9.17). The first term is asymptotically normal:

$$\sqrt{n}\left(\widehat{\theta}-\theta_n\right)\underset{d}{\longrightarrow}\sqrt{V_\theta}Z$$

where $Z\sim \mathrm{N}(0,1)$. Therefore

$$\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\underset{d}{\longrightarrow}\sqrt{V_\theta}Z+h\sim \mathrm{N}(h,V_\theta).$$

This asymptotic distribution depends continuously on the localizing parameter h.

Applied to the t statistic we find

$$T=\frac{\widehat{\theta}-\theta_0}{s(\widehat{\theta})}\underset{d}{\longrightarrow}\frac{\sqrt{V_\theta}Z+h}{\sqrt{V_\theta}}\sim Z+\delta \tag{9.18}$$

where $\delta=h/\sqrt{V_\theta}$. This generalizes Theorem 9.1 (which assumes $\mathbb{H}_0$ is true) to allow for local alternatives of the form (9.17).

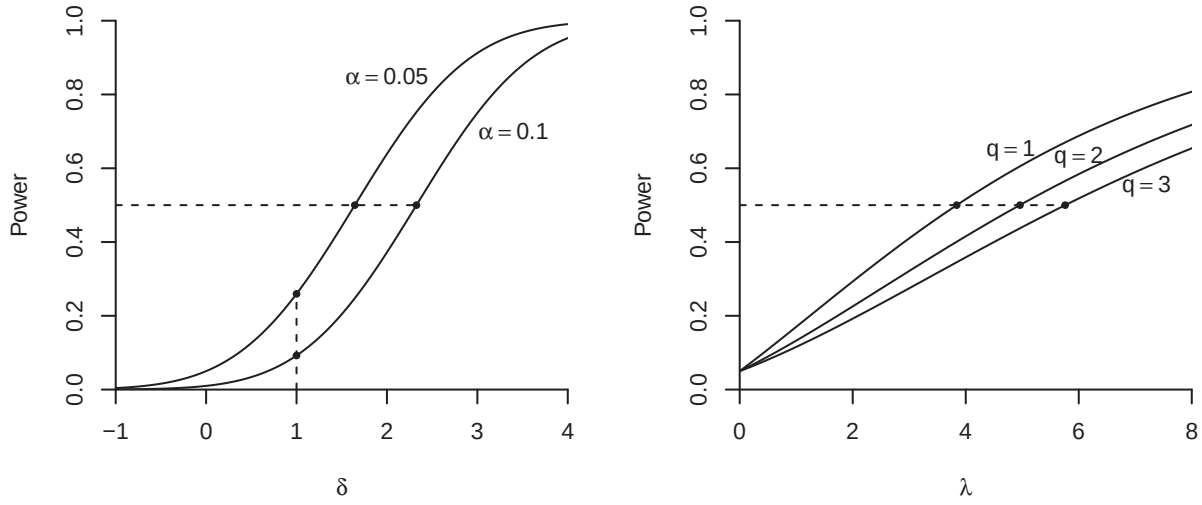
Consider a t-test of $\mathbb{H}_0$ against the one-sided alternative $\mathbb{H}_1:\theta>\theta_0$ which rejects $\mathbb{H}_0$ for $T>c$ where $\Phi(c)=1-\alpha$. The **asymptotic local power** of this test is the limit (as the sample size diverges) of the rejection probability under the local alternative (9.17)

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\left[\text{Reject }\mathbb{H}_0\right]&=\lim_{n\to\infty}\mathbb{P}\left[T>c\right]\\&=\mathbb{P}\left[Z+\delta>c\right]\\&=1-\Phi\left(c-\delta\right)\\&=\Phi\left(\delta-c\right)\\&\overset{\text{def}}{=}\pi(\delta).\end{aligned}$$

We call $\pi(\delta)$ the **asymptotic local power function**.

In Figure 9.3(a) we plot the local power function $\pi(\delta)$ as a function of $\delta\in[-1,4]$ for tests of asymptotic size $\alpha=0.05$ and $\alpha=0.01$. $\delta=0$ corresponds to the null hypothesis so $\pi(\delta)=\alpha$. The power functions are monotonically increasing in δ. Note that the power is lower than α for $\delta<0$ due to the one-sided nature of the test.

We can see that the power functions are ranked by α so that the test with $\alpha=0.05$ 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.



(a) One-Sided t Test

(b) Vector Case

Figure 9.3: Asymptotic Local Power Function

The coefficient δ can be interpreted as the parameter deviation measured as a multiple of the standard error $s(\widehat{\theta})$. To see this, recall that $s(\widehat{\theta}) = n^{-1/2}\sqrt{\widehat{V}_\theta} \simeq n^{-1/2}\sqrt{V_\theta}$ and then note that

$$\delta = \frac{h}{\sqrt{V_\theta}} \simeq \frac{n^{-1/2}h}{s(\widehat{\theta})} = \frac{\theta_n - \theta_0}{s(\widehat{\theta})}.$$

Thus δ approximately equals the deviation $\theta_n - \theta_0$ expressed as multiples of the standard error $s(\widehat{\theta})$. Thus as we examine Figure 9.3(a) 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 4.2 the standard error for the coefficient on "Married Female" is 0.010. Thus, in this example $\delta = 1$ corresponds to $\theta_n = 0.010$ or an 1.0% wage premium for married females. Our calculations show that the asymptotic power of a one-sided 5% test against this alternative is about 26%.

The difference between power functions can be measured either vertically or horizontally. For example, in Figure 9.3(a) there is a vertical dashed line at $\delta = 1$, showing that the asymptotic local power function $\pi(\delta)$ equals 26% for $\alpha = 0.0$, and 9% for $\alpha = 0.01$. This is the difference in power across tests of differing size, holding fixed the parameter in the alternative.

A horizontal comparison can also be illuminating. To illustrate, in Figure 9.3(a) there is a horizontal dashed 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 two power curves at $\delta = 1.65$ ($\alpha = 0.05$) 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, and 2.33 standard errors for a one-sided 1% test.

The ratio of these values (e.g. $2.33/1.65 = 1.41$) measures the relative parameter magnitude needed to achieve the same power. (Thus, for a 1% size test to achieve 50% power, the parameter must be 41% larger than for a 5% size test.) Even more interesting, the square of this ratio (e.g. $1.41^2 = 2$) is the increase in sample size needed to achieve the same power under fixed parameters. That is, to achieve 50% power,

a 1% size test needs twice as many observations as a 5% size test. This interpretation follows by the following informal argument. By definition and (9.17) $\delta = h/\sqrt{V_\theta} = \sqrt{n}(\theta_n - \theta_0)/\sqrt{V_\theta}$. Thus holding θ and V_θ fixed, δ^2 is proportional to n.

The analysis of a two-sided t test is similar. (9.18) implies that

$$T = \left| \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})} \right| \xrightarrow[d]{} |Z + \delta|$$

and thus the local power of a two-sided t test is

$$\lim_{n\to\infty} \mathbb{P}\left[\text{Reject } \mathbb{H}_0\right] = \lim_{n\to\infty} \mathbb{P}[T > c] = \mathbb{P}[|Z + \delta| > c] = \Phi(\delta - c) + \Phi(-\delta - c)$$

which is monotonically increasing in $|\delta|$.

Theorem 9.10 Under Assumptions 7.2, 7.3, 7.4, and $\theta_n = r(\beta_n) = r_0 + n^{-1/2}h$, then

$$T(\theta_0) = \frac{\widehat{\theta} - \theta_0}{s(\widehat{\theta})} \xrightarrow[d]{} Z + \delta$$

where $Z \sim \mathrm{N}(0,1)$ and $\delta = h/\sqrt{V_\theta}$. For c such that $\Phi(c) = 1 - \alpha$,

$$\mathbb{P}[T(\theta_0) > c] \longrightarrow \Phi(\delta - c).$$

Furthermore, for c such that $\Phi(c) = 1 - \alpha/2$,

$$\mathbb{P}[|T(\theta_0)| > c] \longrightarrow \Phi(\delta - c) + \Phi(-\delta - c).$$

9.23 Asymptotic Local Power, Vector Case

In this section we extend the local power analysis of the previous section to the case of vector-valued alternatives. We generalize (9.17) to vector-valued θ_n. The local parameterization is

$$\theta_n = r(\beta_n) = \theta_0 + n^{-1/2}h \tag{9.19}$$

where h is $q \times 1$.

Under (9.19),

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = \sqrt{n}\left(\widehat{\theta} - \theta_n\right) + h \xrightarrow[d]{} Z_h \sim \mathrm{N}(h, \boldsymbol{V}_\theta),$$

a normal random vector with mean h and covariance matrix $\boldsymbol{V}_\theta$.

Applied to the Wald statistic we find

$$W = n\left(\widehat{\theta} - \theta_0\right)' \widehat{\boldsymbol{V}}_\theta^{-1} \left(\widehat{\theta} - \theta_0\right) \xrightarrow[d]{} Z_h' \boldsymbol{V}_\theta^{-1} Z_h \sim \chi_q^2(\lambda) \tag{9.20}$$

where $\lambda = h' \boldsymbol{V}^{-1} h$. $\chi_q^2(\lambda)$ is a non-central chi-square random variable with non-centrality parameter λ. (Theorem 5.3.6.)

The convergence (9.20) shows that under the local alternatives (9.19), $W \xrightarrow[d]{} \chi_q^2(\lambda)$. This generalizes the null asymptotic distribution which obtains as the special case $\lambda = 0$. We can use this result to obtain

a continuous asymptotic approximation to the power function. For any significance level $\alpha > 0$ set the asymptotic critical value c so that $\mathbb{P}\left[\chi^2_q > c\right] = \alpha$. Then as $n \to \infty$,

$$\mathbb{P}[W > c] \longrightarrow \mathbb{P}\left[\chi^2_q(\lambda) > c\right] \stackrel{\text{def}}{=} \pi(\lambda).$$

The asymptotic local power function $\pi(\lambda)$ depends only on α, q, and λ.

Theorem 9.11 Under Assumptions 7.2, 7.3, 7.4, and $\theta_n = r(\beta_n) = \theta_0 + n^{-1/2}h$, then $W \underset{d}{\longrightarrow} \chi^2_q(\lambda)$ where $\lambda = h'\boldsymbol{V}_\theta^{-1}h$. Furthermore, for c such that $\mathbb{P}\left[\chi^2_q > c\right] = \alpha$, $\mathbb{P}[W > c] \longrightarrow \mathbb{P}\left[\chi^2_q(\lambda) > c\right]$.

Figure 9.3(b) plots $\pi(\lambda)$ as a function of λ for $q = 1$, $q = 2$, and $q = 3$, and $\alpha = 0.05$. The asymptotic power functions are monotonically increasing in λ and asymptote to one.

Figure 9.3(b) 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 dashed line at 50% power. The dashed 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 maintain 50% power.

9.24 Exercises

Exercise 9.1 Prove that if an additional regressor $\boldsymbol{X}_{k+1}$ is added to $\boldsymbol{X}$, Theil's adjusted $\overline{R}^2$ increases if and only if $|T_{k+1}| > 1$, where $T_{k+1} = \widehat{\beta}_{k+1}/s(\widehat{\beta}_{k+1})$ is the t-ratio for $\widehat{\beta}_{k+1}$ and

$$s(\widehat{\beta}_{k+1}) = \left(s^2[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{k+1,k+1}\right)^{1/2}$$

is the homoskedasticity-formula standard error.

Exercise 9.2 You have two independent samples (Y_{1i}, X_{1i}) and (Y_{2i}, X_{2i}) both with sample sizes n which satisfy $Y_1 = X_1'\beta_1 + e_1$ and $Y_2 = X_2'\beta_2 + e_2$, where $\mathbb{E}[X_1 e_1] = 0$ and $\mathbb{E}[X_2 e_2] = 0$. Let $\widehat{\beta}_1$ and $\widehat{\beta}_2$ be the OLS estimators of $\beta_1 \in \mathbb{R}^k$ and $\beta_2 \in \mathbb{R}^k$.

(a) Find the asymptotic distribution of $\sqrt{n}\left(\left(\widehat{\beta}_2 - \widehat{\beta}_1\right) - \left(\beta_2 - \beta_1\right)\right)$ as $n \to \infty$.

(b) Find an appropriate test statistic for $\mathbb{H}_0 : \beta_2 = \beta_1$.

(c) Find the asymptotic distribution of this statistic under $\mathbb{H}_0$.

Exercise 9.3 Let T be a t-statistic for $\mathbb{H}_0 : \theta = 0$ versus $\mathbb{H}_1 : \theta \neq 0$. Since $|T| \to_d |Z|$ under $\mathbb{H}_0$, someone suggests the test "Reject $\mathbb{H}_0$ if $|T| < c_1$ or $|T| > c_2$, where c_1 is the $\alpha/2$ quantile of $|Z|$ and c_2 is the $1 - \alpha/2$ quantile of $|Z|$.

(a) Show that the asymptotic size of the test is α.

(b) Is this a good test of $\mathbb{H}_0$ versus $\mathbb{H}_1$? Why or why not?

Exercise 9.4 Let W be a Wald statistic for $\mathbb{H}_0 : \theta = 0$ versus $\mathbb{H}_1 : \theta \neq 0$, where θ is $q \times 1$. Since $W \underset{d}{\longrightarrow} \chi^2_q$ under H_0, someone suggests the test "Reject $\mathbb{H}_0$ if $W < c_1$ or $W > c_2$, where c_1 is the $\alpha/2$ quantile of χ^2_q and c_2 is the $1-\alpha/2$ quantile of χ^2_q.

(a) Show that the asymptotic size of the test is α.

(b) Is this a good test of $\mathbb{H}_0$ versus $\mathbb{H}_1$? Why or why not?

Exercise 9.5 Take the linear model $Y = X_1'\beta_1 + X_2'\beta_2 + e$ with $\mathbb{E}[Xe] = 0$ where both X_1 and X_2 are $q \times 1$. Show how to test the hypotheses $\mathbb{H}_0 : \beta_1 = \beta_2$ against $\mathbb{H}_1 : \beta_1 \neq \beta_2$.

Exercise 9.6 Suppose a researcher wants to know which of a set of 20 regressors has an effect on a variable *testscore*. He regresses *testscore* on the 20 regressors and reports the results. One of the 20 regressors (*studytime*) has a large t-ratio (about 2.5), while the other t-ratios are insignificant (smaller than 2 in absolute value). He argues that the data show that *studytime* is the key predictor for *testscore*. Do you agree with this conclusion? Is there a deficiency in his reasoning?

Exercise 9.7 Take the model $Y = X\beta_1 + X^2\beta_2 + e$ with $\mathbb{E}[e \mid X] = 0$ where Y is wages (dollars per hour) and X is age. Describe how you would test the hypothesis that the expected wage for a 40-year-old worker is $20 an hour.

Exercise 9.8 You want to test $\mathbb{H}_0 : \beta_2 = 0$ against $\mathbb{H}_1 : \beta_2 \neq 0$ in the model $Y = X_1'\beta_1 + X_2'\beta_2 + e$ with $\mathbb{E}[Xe] = 0$. You read a paper which estimates the model

$$Y = X_1'\widehat{\gamma}_1 + (X_2 - X_1)'\widehat{\gamma}_2 + u$$

and reports a test of $\mathbb{H}_0 : \gamma_2 = 0$ against $\mathbb{H}_1 : \gamma_2 \neq 0$. Is this related to the test you wanted to conduct?

Exercise 9.9 Suppose a researcher uses one dataset to test a specific hypothesis $\mathbb{H}_0$ against $\mathbb{H}_1$ and finds that he can reject $\mathbb{H}_0$. A second researcher gathers a similar but independent dataset, uses similar methods and finds that she cannot reject $\mathbb{H}_0$. How should we (as interested professionals) interpret these mixed results?

Exercise 9.10 In Exercise 7.8 you showed that $\sqrt{n}\left(\widehat{\sigma}^2 - \sigma^2\right) \underset{d}{\longrightarrow} \mathrm{N}(0, V)$ as $n \to \infty$ for some V. Let $\widehat{V}$ be an estimator of V.

(a) Using this result construct a t-statistic for $\mathbb{H}_0 : \sigma^2 = 1$ against $\mathbb{H}_1 : \sigma^2 \neq 1$.

(b) Using the Delta Method find the asymptotic distribution of $\sqrt{n}(\widehat{\sigma} - \sigma)$.

(c) Use the previous result to construct a t-statistic for $\mathbb{H}_0 : \sigma = 1$ against $\mathbb{H}_1 : \sigma \neq 1$.

(d) Are the null hypotheses in (a) and (c) the same or are they different? Are the tests in (a) and (c) the same or are they different? If they are different, describe a context in which the two tests would give contradictory results.

Exercise 9.11 Consider a regression such as Table 4.1 where both *experience* and its square are included. A researcher wants to test the hypothesis that *experience* does not affect mean wages and does this by computing the t-statistic for *experience*. Is this the correct approach? If not, what is the appropriate testing method?

Exercise 9.12 A researcher estimates a regression and computes a test of $\mathbb{H}_0$ against $\mathbb{H}_1$ and finds a p-value of $p = 0.08$, or "not significant". She says "I need more data. If I had a larger sample the test will have more power and then the test will reject." Is this interpretation correct?

Exercise 9.13 A common view is that "If the sample size is large enough, any hypothesis will be rejected." What does this mean? Interpret and comment.

Exercise 9.14 Take the model $Y = X'\beta + e$ with $\mathbb{E}[Xe] = 0$ and parameter of interest $\theta = \boldsymbol{R}'\beta$ with $\boldsymbol{R}$ $k \times 1$. Let $\widehat{\beta}$ be the least squares estimator and $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$ its variance estimator.

(a) Write down $\widehat{C}$, the 95% asymptotic confidence interval for θ, in terms of $\widehat{\beta}$, $\widehat{\boldsymbol{V}}_{\widehat{\beta}}$, $\boldsymbol{R}$, and $z = 1.96$ (the 97.5% quantile of $\mathrm{N}(0,1)$).

(b) Show that the decision "Reject $\mathbb{H}_0$ if $\theta_0 \notin \widehat{C}$" is an asymptotic 5% test of $\mathbb{H}_0 : \theta = \theta_0$.

Exercise 9.15 You are at a seminar where a colleague presents a simulation study of a test of a hypothesis $\mathbb{H}_0$ with nominal size 5%. Based on $B = 100$ simulation replications under $\mathbb{H}_0$ the estimated size is 7%. Your colleague says: "Unfortunately the test over-rejects."

(a) Do you agree or disagree with your colleague? Explain. Hint: Use an asymptotic (large B) approximation.

(b) Suppose the number of simulation replications were $B = 1000$ yet the estimated size is still 7%. Does your answer change?

Exercise 9.16 Consider two alternative regression models

$$\begin{aligned} Y &= X_1'\beta_1 + e_1 \\ \mathbb{E}[X_1 e_1] &= 0 \end{aligned} \tag{9.21}$$

$$\begin{aligned} Y &= X_2'\beta_2 + e_2 \\ \mathbb{E}[X_2 e_2] &= 0 \end{aligned} \tag{9.22}$$

where X_1 and X_2 have at least some different regressors. (For example, (9.21) is a wage regression on geographic variables and (2) is a wage regression on personal appearance measurements.) You want to know if model (9.21) or model (9.22) fits the data better. Define $\sigma_1^2 = \mathbb{E}\left[e_1^2\right]$ and $\sigma_2^2 = \mathbb{E}\left[e_2^2\right]$. You decide that the model with the smaller variance fit (e.g., model (9.21) fits better if $\sigma_1^2 < \sigma_2^2$.) You decide to test for this by testing the hypothesis of equal fit $\mathbb{H}_0 : \sigma_1^2 = \sigma_2^2$ against the alternative of unequal fit $\mathbb{H}_1 : \sigma_1^2 \neq \sigma_2^2$. For simplicity, suppose that e_{1i} and e_{2i} are observed.

(a) Construct an estimator $\widehat{\theta}$ of $\theta = \sigma_1^2 - \sigma_2^2$.

(b) Find the asymptotic distribution of $\sqrt{n}\left(\widehat{\theta} - \theta\right)$ as $n \to \infty$.

(c) Find an estimator of the asymptotic variance of $\widehat{\theta}$.

(d) Propose a test of asymptotic size α of $\mathbb{H}_0$ against $\mathbb{H}_1$.

(e) Suppose the test accepts $\mathbb{H}_0$. Briefly, what is your interpretation?

Exercise 9.17 You have two regressors X_1 and X_2 and estimate a regression with all quadratic terms included

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 + e.$$

One of your advisors asks: Can we exclude the variable X_2 from this regression?

How do you translate this question into a statistical test? When answering these questions, be specific, not general.

(a) What is the relevant null and alternative hypotheses?

(b) What is an appropriate test statistic?

(c) What is the appropriate asymptotic distribution for the statistic?

(d) What is the rule for acceptance/rejection of the null hypothesis?

Exercise 9.18 The observed data is $\{Y_i, X_i, Z_i\} \in \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}^\ell$, $k > 1$ and $\ell > 1$, $i = 1, ..., n$. An econometrician first estimates $Y_i = X_i' \widehat{\beta} + \widehat{e}_i$ by least squares. The econometrician next regresses the residual $\widehat{e}_i$ on Z_i, which can be written as $\widehat{e}_i = Z_i' \widetilde{\gamma} + \widetilde{u}_i$.

(a) Define the population parameter γ being estimated in this second regression.

(b) Find the probability limit for $\widetilde{\gamma}$.

(c) Suppose the econometrician constructs a Wald statistic W for $\mathbb{H}_0 : \gamma = 0$ from the second regression, ignoring the two-stage estimation process. Write down the formula for W.

(d) Assume $\mathbb{E}\left[ZX'\right] = 0$. Find the asymptotic distribution for W under $\mathbb{H}_0 : \gamma = 0$.

(e) If $\mathbb{E}\left[ZX'\right] \neq 0$ will your answer to (d) change?

Exercise 9.19 An economist estimates $Y = X_1' \beta_1 + X_2 \beta_2 + e$ by least squares and tests the hypothesis $\mathbb{H}_0 : \beta_2 = 0$ against $\mathbb{H}_1 : \beta_2 \neq 0$. Assume $\beta_1 \in \mathbb{R}^k$ and $\beta_2 \in \mathbb{R}$. She obtains a Wald statistic $W = 0.34$. The sample size is $n = 500$.

(a) What is the correct degrees of freedom for the χ^2 distribution to evaluate the significance of the Wald statistic?

(b) The Wald statistic W is very small. Indeed, is it less than the 1% quantile of the appropriate χ^2 distribution? If so, should you reject $\mathbb{H}_0$? Explain your reasoning.

Exercise 9.20 You are reading a paper, and it reports the results from two nested OLS regressions:

$$Y_i = X_{1i}' \widetilde{\beta}_1 + \widetilde{e}_i$$
$$Y_i = X_{1i}' \widehat{\beta}_1 + X_{2i}' \widehat{\beta}_2 + \widehat{e}_i.$$

Some summary statistics are reported:

Short Regression	Long Regression
$R^2 = .20$	$R^2 = .26$
$\sum_{i=1}^n \widetilde{e}_i^2 = 106$	$\sum_{i=1}^n \widehat{e}_i^2 = 100$
# of coefficients=5	# of coefficients=8
$n = 50$	$n = 50$

You are curious if the estimate $\widehat{\beta}_2$ is statistically different from the zero vector. Is there a way to determine an answer from this information? Do you have to make any assumptions (beyond the standard regularity conditions) to justify your answer?

Exercise 9.21 Take the model $Y = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + e$ with $\mathbb{E}[Xe] = 0$. Describe how to test

$$\mathbb{H}_0 : \frac{\beta_1}{\beta_2} = \frac{\beta_3}{\beta_4}$$

against

$$\mathbb{H}_1 : \frac{\beta_1}{\beta_2} \neq \frac{\beta_3}{\beta_4}.$$

Exercise 9.22 You have a random sample from the model $Y = X\beta_1 + X^2\beta_2 + e$ with $\mathbb{E}[e \mid X] = 0$ where Y is wages (dollars per hour) and X is age. Describe how you would test the hypothesis that the expected wage for a 40-year-old worker is $20 an hour.

Exercise 9.23 Let T be a test statistic such that under $\mathbb{H}_0$, $T \underset{d}{\longrightarrow} \chi^2_3$. Since $\mathbb{P}\left[\chi^2_3 > 7.815\right] = 0.05$, an asymptotic 5% test of $\mathbb{H}_0$ rejects when $T > 7.815$. An econometrician is interested in the Type I error of this test when $n = 100$ and the data structure is well specified. She performs the following Monte Carlo experiment.

- $B = 200$ samples of size $n = 100$ are generated from a distribution satisfying $\mathbb{H}_0$.
- On each sample, the test statistic T_b is calculated.
- She calculates $\widehat{p} = B^{-1}\sum_{b=1}^{B} \mathbb{1}\{T_b > 7.815\} = 0.070$.
- The econometrician concludes that the test T is oversized in this context – it rejects too frequently under $\mathbb{H}_0$.

Is her conclusion correct, incorrect, or incomplete? Be specific in your answer.

Exercise 9.24 Do a Monte Carlo simulation. Take the model $Y = \alpha + X\beta + e$ with $\mathbb{E}[Xe] = 0$ where the parameter of interest is $\theta = \exp(\beta)$. Your data generating process (DGP) for the simulation is: X is $U[0,1]$, $e \sim \mathrm{N}(0,1)$ is independent of X, and $n = 50$. Set $\alpha = 0$ and $\beta = 1$. Generate $B = 1000$ independent samples with α. On each, estimate the regression by least squares, calculate the covariance matrix using a standard (heteroskedasticity-robust) formula, and similarly estimate θ and its standard error. For each replication, store $\widehat{\beta}$, $\widehat{\theta}$, $T_\beta = \left(\widehat{\beta} - \beta\right)/s\left(\widehat{\beta}\right)$, and $T_\theta = \left(\widehat{\theta} - \theta\right)/s\left(\widehat{\theta}\right)$.

(a) Does the value of α matter? Explain why the described statistics are **invariant** to α and thus setting $\alpha = 0$ is irrelevant.

(b) From the 1000 replications estimate $\mathbb{E}\left[\widehat{\beta}\right]$ and $\mathbb{E}\left[\widehat{\theta}\right]$. Discuss if you see evidence if either estimator is biased or unbiased.

(c) From the 1000 replications estimate $\mathbb{P}\left[T_\beta > 1.645\right]$ and $\mathbb{P}[T_\theta > 1.645]$. What does asymptotic theory predict these probabilities should be in large samples? What do your simulation results indicate?

Exercise 9.25 The data set `Invest1993` on the textbook website contains data on 1962 U.S. firms extracted from Compustat, assembled by Bronwyn Hall, and used in Hall and Hall (1993).

The variables we use in this exercise are in the table below. The flow variables are annual sums. The stock variables are beginning of year.

	year	year of the observation
I	inva	Investment to Capital Ratio
Q	vala	Total Market Value to Asset Ratio (Tobin's Q)
C	cfa	Cash Flow to Asset Ratio
D	debta	Long Term Debt to Asset Ratio

(a) Extract the sub-sample of observations for 1987. There should be 1028 observations. Estimate a linear regression of I (investment to capital ratio) 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 (Tobin's Q). This theory predicts that the coefficient on Q should be positive and the others should be zero. Test the joint hypothesis that the coefficients on cash flow (C) and debt (D) are zero. Test the hypothesis that the coefficient on Q is zero. Are the results consistent with the predictions of the theory?

(d) Now try a nonlinear (quadratic) specification. Regress I on Q, C, D, Q^2, C^2, D^2, $Q \times C$, $Q \times D$, $C \times D$. Test the joint hypothesis that the six interaction and quadratic coefficients are zero.

Exercise 9.26 In a paper in 1963, Marc Nerlove analyzed a cost function for 145 American electric companies. Nerlov was interested in estimating a *cost function*: $C = f(Q, PL, PF, PK)$ where the variables are listed in the table below. His data set `Nerlove1963` is on the textbook website.

C	Total Cost
Q	Output
PL	Unit price of labor
PK	Unit price of capital
PF	Unit price of fuel

(a) First, estimate an unrestricted Cobb-Douglass specification

$$\log C = \beta_1 + \beta_2 \log Q + \beta_3 \log PL + \beta_4 \log PK + \beta_5 \log PF + e. \tag{9.23}$$

Report parameter estimates and standard errors.

(b) What is the economic meaning of the restriction $\mathbb{H}_0 : \beta_3 + \beta_4 + \beta_5 = 1$?

(c) Estimate (9.23) by constrained least squares imposing $\beta_3 + \beta_4 + \beta_5 = 1$. Report your parameter estimates and standard errors.

(d) Estimate (9.23) by efficient minimum distance imposing $\beta_3 + \beta_4 + \beta_5 = 1$. Report your parameter estimates and standard errors.

(e) Test $\mathbb{H}_0 : \beta_3 + \beta_4 + \beta_5 = 1$ using a Wald statistic.

(f) Test $\mathbb{H}_0 : \beta_3 + \beta_4 + \beta_5 = 1$ using a minimum distance statistic.

Exercise 9.27 In Section 8.12 we reported estimates from Mankiw, Romer and Weil (1992). We reported estimation both by unrestricted least squares and by constrained estimation, imposing the constraint that three coefficients (2^{nd}, 3^{rd} and 4^{th} coefficients) sum to zero as implied by the Solow growth theory. Using the same dataset `MRW1992` estimate the unrestricted model and test the hypothesis that the three coefficients sum to zero.

Exercise 9.28 Using the `cps09mar` dataset and the subsample of non-Hispanic Black individuals (race code = 2) test the hypothesis that marriage status does not affect mean wages.

(a) Take the regression reported in Table 4.1. Which variables will need to be omitted to estimate a regression for this subsample?

(b) Express the hypothesis "marriage status does not affect mean wages" as a restriction on the coefficients. How many restrictions is this?

(c) Find the Wald (or F) statistic for this hypothesis. What is the appropriate distribution for the test statistic? Calculate the p-value of the test.

(d) What do you conclude?

Exercise 9.29 Using the `cps09mar` dataset and the subsample of non-Hispanic Black individuals (race code = 2) and white individuals (race code = 1) test the hypothesis that the returns to education is common across groups.

(a) Allow the return to education to vary across the four groups (white male, white female, Black male, Black female) by interacting dummy variables with education. Estimate an appropriate version of the regression reported in Table 4.1.

(b) Find the Wald (or F) statistic for this hypothessis. What is the appropriate distribution for the test statistic? Calculate the p-value of the test.

(c) What do you conclude?

Chapter 10

Resampling Methods

10.1 Introduction

So far in this textbook we have discussed two approaches to inference: exact and asymptotic. Both have their strengths and weaknesses. Exact theory provides a useful benchmark but is based on the unrealistic and stringent assumption of the homoskedastic normal regression model. Asymptotic theory provides a more flexible distribution theory but is an approximation with uncertain accuracy.

In this chapter we introduce a set of alternative inference methods which are based around the concept of resampling – which means using sampling information extracted from the empirical distribution of the data. These are powerful methods, widely applicable, and often more accurate than exact methods and asymptotic approximations. Two disadvantages, however, are (1) resampling methods typically require more computation power; and (2) the theory is considerably more challenging. A consequence of the computation requirement is that most empirical researchers use asymptotic approximations for routine calculations while resampling approximations are used for final reporting.

We will discuss two categories of resampling methods used in statistical and econometric practice: jackknife and bootstrap. Most of our attention will be given to the bootstrap as it is the most commonly used resampling method in econometric practice.

The **jackknife** is the distribution obtained from the n leave-one-out estimators (see Section 3.20). The jackknife is most commonly used for variance estimation.

The **bootstrap** is the distribution obtained by estimation on samples created by i.i.d. sampling with replacement from the dataset. (There are other variants of bootstrap sampling, including parametric sampling and residual sampling.) The bootstrap is commonly used for variance estimation, confidence interval construction, and hypothesis testing.

There is a third category of resampling methods known as **sub-sampling** which we will not cover in this textbook. Sub-sampling is the distribution obtained by estimation on sub-samples (sampling without replacement) of the dataset. Sub-sampling can be used for most of same purposes as the bootstrap. See the excellent monograph by Politis, Romano and Wolf (1999).

10.2 Example

To motivate our discussion we focus on the application presented in Section 3.7, which is a bivariate regression applied to the CPS subsample of married Black female wage earners with 12 years potential work experience and displayed in Table 3.1. The regression equation is

$$\log(wage) = \beta_1 education + \beta_2 + e.$$

The estimates as reported in (4.44) are

$$\begin{aligned} \log(wage) = & \underset{(0.031)}{0.155} \quad education + \underset{(0.493)}{0.698} \quad + \widehat{e} \\ \widehat{\sigma}^2 = & \underset{(0.043)}{0.144} \\ n = & \, 20. \end{aligned}$$

We focus on four estimates constructed from this regression. The first two are the coefficient estimates $\widehat{\beta}_1$ and $\widehat{\beta}_2$. The third is the variance estimate $\widehat{\sigma}^2$. The fourth is an estimate of the expected level of wages for an individual with 16 years of education (a college graduate), which turns out to be a nonlinear function of the parameters. Under the simplifying assumption that the error e is independent of the level of education and normally distributed we find that the expected level of wages is

$$\begin{aligned} \mu &= \mathbb{E}\left[wage \mid education = 16\right] \\ &= \mathbb{E}\left[\exp\left(16\beta_1 + \beta_2 + e\right)\right] \\ &= \exp\left(16\beta_1 + \beta_2\right)\mathbb{E}\left[\exp(e)\right] \\ &= \exp\left(16\beta_1 + \beta_2 + \sigma^2/2\right). \end{aligned}$$

The final equality is $\mathbb{E}\left[\exp(e)\right] = \exp\left(\sigma^2/2\right)$ which can be obtained from the normal moment generating function. The parameter μ is a nonlinear function of the coefficients. The natural estimator of μ replaces the unknowns by the point estimators. Thus

$$\widehat{\mu} = \exp\left(16\widehat{\beta}_1 + \widehat{\beta}_2 + \widehat{\sigma}^2/2\right) = \underset{(2.29)}{25.80}$$

The standard error for $\widehat{\mu}$ can be found by extending Exercise 7.8 to find the joint asymptotic distribution of $\widehat{\sigma}^2$ and the slope estimates, and then applying the delta method.

We are interested in calculating standard errors and confidence intervals for the four estimates described above.

10.3 Jackknife Estimation of Variance

The jackknife estimates moments of estimators using the distribution of the leave-one-out estimators. The jackknife estimators of bias and variance were introduced by Quenouille (1949) and Tukey (1958), respectively. The idea was expanded further in the monographs of Efron (1982) and Shao and Tu (1995).

Let $\widehat{\theta}$ be any estimator of a vector-valued parameter θ which is a function of a random sample of size n. Let $\boldsymbol{V}_{\widehat{\theta}} = \text{var}\left[\widehat{\theta}\right]$ be the variance of $\widehat{\theta}$. Define the leave-one-out estimators $\widehat{\theta}_{(-i)}$ which are computed using the formula for $\widehat{\theta}$ except that observation i is deleted. Tukey's jackknife estimator for $\boldsymbol{V}_{\widehat{\theta}}$ is defined as a scale of the sample variance of the leave-one-out estimators:

$$\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{\text{jack}} = \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{(-i)} - \overline{\theta}\right)\left(\widehat{\theta}_{(-i)} - \overline{\theta}\right)' \tag{10.1}$$

where $\overline{\theta}$ is the sample mean of the leave-one-out estimators $\overline{\theta} = n^{-1}\sum_{i=1}^{n}\widehat{\theta}_{(-i)}$. For scalar estimators $\widehat{\theta}$ the jackknife standard error is the square root of (10.1): $s_{\widehat{\theta}}^{\text{jack}} = \sqrt{\widehat{V}_{\widehat{\theta}}^{\text{jack}}}$.

A convenient feature of the jackknife estimator $\widehat{V}_{\widehat{\theta}}^{\text{jack}}$ is that the formula (10.1) is quite general and does not require any technical (exact or asymptotic) calculations. A downside is that can require n separate estimations, which in some cases can be computationally costly.

In most cases $\widehat{V}_{\widehat{\theta}}^{\text{jack}}$ will be similar to a robust asymptotic covariance matrix estimator. The main attractions of the jackknife estimator are that it can be used when an explicit asymptotic variance formula is not available and that it can be used as a check on the reliability of an asymptotic formula.

The formula (10.1) is not immediately intuitive so may benefit from some motivation. We start by examining the sample mean $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ for $Y \in \mathbb{R}^m$. The leave-one-out estimator is

$$\overline{Y}_{(-i)} = \frac{1}{n-1}\sum_{j\neq i} Y_j = \frac{n}{n-1}\overline{Y} - \frac{1}{n-1}Y_i. \tag{10.2}$$

The sample mean of the leave-one-out estimators is

$$\frac{1}{n}\sum_{i=1}^{n}\overline{Y}_{(-i)} = \frac{n}{n-1}\overline{Y} - \frac{1}{n-1}\overline{Y} = \overline{Y}.$$

The difference is

$$\overline{Y}_{(-i)} - \overline{Y} = \frac{1}{n-1}\left(\overline{Y} - Y_i\right).$$

The jackknife estimate of variance (10.1) is then

$$\begin{aligned}\widehat{V}_{\overline{Y}}^{\text{jack}} &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\frac{1}{n-1}\right)^2\left(\overline{Y} - Y_i\right)\left(\overline{Y} - Y_i\right)' \\ &= \frac{1}{n}\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\overline{Y} - Y_i\right)\left(\overline{Y} - Y_i\right)'. \end{aligned} \tag{10.3}$$

This is identical to the conventional estimator for the variance of $\overline{Y}$. Indeed, Tukey proposed the $(n-1)/n$ scaling in (10.1) so that $\widehat{V}_{\overline{Y}}^{\text{jack}}$ precisely equals the conventional estimator.

We next examine the case of least squares regression coefficient estimator. Recall from (3.43) that the leave-one-out OLS estimator equals

$$\widehat{\beta}_{(-i)} = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \widetilde{e}_i \tag{10.4}$$

where $\widetilde{e}_i = (1-h_{ii})^{-1}\widehat{e}_i$ and $h_{ii} = X_i'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}X_i$. The sample mean of the leave-one-out estimators is $\overline{\beta} = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\widetilde{\mu}$ where $\widetilde{\mu} = n^{-1}\sum_{i=1}^{n} X_i\widetilde{e}_i$. Thus $\widehat{\beta}_{(-i)} - \overline{\beta} = -\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(X_i\widetilde{e}_i - \widetilde{\mu}\right)$. The jackknife estimate of variance for $\widehat{\beta}$ is

$$\begin{aligned}\widehat{V}_{\widehat{\beta}}^{\text{jack}} &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\beta}_{(-i)} - \overline{\beta}\right)\left(\widehat{\beta}_{(-i)} - \overline{\beta}\right)' \\ &= \frac{n-1}{n}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i'\widetilde{e}_i^2 - n\widetilde{\mu}\widetilde{\mu}'\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \frac{n-1}{n}\widehat{V}_{\widehat{\beta}}^{\text{HC3}} - (n-1)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\widetilde{\mu}\widetilde{\mu}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \end{aligned} \tag{10.5}$$

where $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC3}}$ is the HC3 covariance estimator (4.39) based on prediction errors. The second term in (10.5) is typically quite small since $\widetilde{\mu}$ is typically small in magnitude. Thus $\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{jack}} \simeq \widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{HC3}}$. Indeed the HC3 estimator was originally motivated as a simplification of the jackknife estimator. This shows that for regression coefficients the jackknife estimator of variance is similar to a conventional robust estimator. This is accomplished without the user "knowing" the form of the asymptotic covariance matrix. This is further confirmation that the jackknife is making a reasonable calculation.

Third, we examine the jackknife estimator for a function $\widehat{\theta} = r(\widehat{\beta})$ of a least squares estimator. The leave-one-out estimator of θ is

$$\begin{aligned}\widehat{\theta}_{(-i)} &= r(\widehat{\beta}_{(-i)}) \\ &= r\left(\widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \widetilde{e}_i\right) \\ &\simeq \widehat{\theta} - \widehat{\boldsymbol{R}}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} X_i \widetilde{e}_i.\end{aligned}$$

The second equality is (10.4). The final approximation is obtained by a mean-value expansion, using $r(\widehat{\beta}) = \widehat{\theta}$ and setting $\widehat{\boldsymbol{R}} = \left(\partial/\partial\beta\right) r(\widehat{\beta})'$. This approximation holds in large samples because $\widehat{\beta}_{(-i)}$ are uniformly consistent for β. The jackknife variance estimator for $\widehat{\theta}$ thus equals

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{\mathrm{jack}} &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{(-i)} - \overline{\theta}\right)\left(\widehat{\theta}_{(-i)} - \overline{\theta}\right)' \\ &\simeq \frac{n-1}{n}\widehat{\boldsymbol{R}}'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widetilde{e}_i^2 - n\widetilde{\mu}\widetilde{\mu}'\right)\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\widehat{\boldsymbol{R}} \\ &= \widehat{\boldsymbol{R}}'\widehat{\boldsymbol{V}}_{\widehat{\beta}}^{\mathrm{jack}}\widehat{\boldsymbol{R}} \\ &\simeq \widehat{\boldsymbol{R}}'\widetilde{\boldsymbol{V}}_{\widehat{\beta}}\widehat{\boldsymbol{R}}.\end{aligned}$$

The final line equals a delta-method estimator for the variance of $\widehat{\theta}$ constructed with the covariance estimator (4.39). This shows that the jackknife estimator of variance for $\widehat{\theta}$ is approximately an asymptotic delta-method estimator. While this is an asymptotic approximation, it again shows that the jackknife produces an estimator which is asymptotically similar to one produced by asymptotic methods. This is despite the fact that the jackknife estimator is calculated without reference to asymptotic theory and does not require calculation of the derivatives of $r(\beta)$.

This argument extends directly to any "smooth function" estimator. Most of the estimators discussed so far in this textbook take the form $\widehat{\theta} = g\left(\overline{W}\right)$ where $\overline{W} = n^{-1}\sum_{i=1}^{n} W_i$ and W_i is some vector-valued function of the data. For any such estimator $\widehat{\theta}$ the leave-one-out estimator equals $\widehat{\theta}_{(-i)} = g\left(\overline{W}_{(-i)}\right)$ and its jackknife estimator of variance is (10.1). Using (10.2) and a mean-value expansion we have the large-sample approximation

$$\begin{aligned}\widehat{\theta}_{(-i)} &= g\left(\overline{W}_{(-i)}\right) \\ &= g\left(\frac{n}{n-1}\overline{W} - \frac{1}{n-1}W_i\right) \\ &\simeq g\left(\overline{W}\right) - \frac{1}{n-1}\boldsymbol{G}\left(\overline{W}\right)' W_i\end{aligned}$$

where $\boldsymbol{G}(x) = (\partial/\partial x)\, g(x)'$. Thus

$$\widehat{\theta}_{(-i)} - \overline{\theta} \simeq -\frac{1}{n-1}\boldsymbol{G}\left(\overline{W}\right)'\left(W_i - \overline{W}\right)$$

and the jackknife estimator of the variance of $\widehat{\theta}$ approximately equals

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{\text{jack}} &= \frac{n-1}{n}\sum_{i=1}^{n}\left(\widehat{\theta}_{(-i)}-\widehat{\theta}_{(\cdot)}\right)\left(\widehat{\theta}_{(-i)}-\widehat{\theta}_{(\cdot)}\right)' \\ &\simeq \frac{n-1}{n}\boldsymbol{G}\left(\overline{W}\right)'\left(\frac{1}{(n-1)^2}\sum_{i=1}^{n}\left(W_i-\overline{W}\right)\left(W_i-\overline{W}\right)'\right)\boldsymbol{G}\left(\overline{W}\right) \\ &= \boldsymbol{G}\left(\overline{W}\right)'\widehat{\boldsymbol{V}}_{\overline{W}}^{\text{jack}}\boldsymbol{G}\left(\overline{W}\right)\end{aligned}$$

where $\widehat{\boldsymbol{V}}_{\overline{W}}^{\text{jack}}$ as defined in (10.3) is the conventional (and jackknife) estimator for the variance of $\overline{W}$. Thus $\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{\text{jack}}$ is approximately the delta-method estimator. Once again, we see that the jackknife estimator automatically calculates what is effectively the delta-method variance estimator, but without requiring the user to explicitly calculate the derivative of $g(x)$.

10.4 Example

We illustrate by reporting the asymptotic and jackknife standard errors for the four parameter estimates given earlier. In Table 10.1 we report the actual values of the leave-one-out estimates for each of the twenty observations in the sample. The jackknife standard errors are calculated as the scaled square roots of the sample variances of these leave-one-out estimates and are reported in the second-to-last row. For comparison the asymptotic standard errors are reported in the final row.

For all estimates the jackknife and asymptotic standard errors are quite similar. This reinforces the credibility of both standard error estimates. The largest differences arise for $\widehat{\beta}_2$ and $\widehat{\mu}$, whose jackknife standard errors are about 5% larger than the asymptotic standard errors.

The take-away from our presentation is that the jackknife is a simple and flexible method for variance and standard error calculation. Circumventing technical asymptotic and exact calculations, the jackknife produces estimates which in many cases are similar to asymptotic delta-method counterparts. The jackknife is especially appealing in cases where asymptotic standard errors are not available or are difficult to calculate. They can also be used as a double-check on the reasonableness of asymptotic delta-method calculations.

In Stata, jackknife standard errors for coefficient estimates in many models are obtained by the `vce(jackknife)` option. For nonlinear functions of the coefficients or other estimators the `jackknife` command can be combined with any other command to obtain jackknife standard errors.

To illustrate, below we list the Stata commands which calculate the jackknife standard errors listed above. The first line is least squares estimation with standard errors calculated by the jackknife. The second line calculates the error variance estimate $\widehat{\sigma}^2$ with a jackknife standard error. The third line does the same for the estimate $\widehat{\mu}$.

Stata Commands

```
reg wage education if mbf12 == 1, vce(jackknife)
jackknife (e(rss)/e(N)): reg wage education if mbf12 == 1
jackknife exp(16*_b[education]+_b[_cons]+e(rss)/e(N)/2): ///
    reg wage education if mbf12 == 1
```

Table 10.1: Leave-one-out Estimators and Jackknife Standard Errors

Observation	$\widehat{\beta}_{1(-i)}$	$\widehat{\beta}_{2(-i)}$	$\widehat{\sigma}^2_{(-i)}$	$\widehat{\mu}_{(-i)}$
1	0.150	0.764	0.150	25.63
2	0.148	0.798	0.149	25.48
3	0.153	0.739	0.151	25.97
4	0.156	0.695	0.144	26.31
5	0.154	0.701	0.146	25.38
6	0.158	0.655	0.151	26.05
7	0.152	0.705	0.114	24.32
8	0.146	0.822	0.147	25.37
9	0.162	0.588	0.151	25.75
10	0.157	0.693	0.139	26.40
11	0.168	0.510	0.141	26.40
12	0.158	0.691	0.118	26.48
13	0.139	0.974	0.141	26.56
14	0.169	0.451	0.131	26.26
15	0.146	0.852	0.150	24.93
16	0.156	0.696	0.148	26.06
17	0.165	0.513	0.140	25.22
18	0.155	0.698	0.151	25.90
19	0.152	0.742	0.151	25.73
20	0.155	0.697	0.151	25.95
s^{jack}	0.032	0.514	0.046	2.39
s^{asy}	0.031	0.493	0.043	2.29

10.5 Jackknife for Clustered Observations

In Section 4.21 we introduced the clustered regression model, cluster-robust variance estimators, and cluster-robust standard errors. Jackknife variance estimation can also be used for clustered samples but with some natural modifications. Recall that the least squares estimator in the clustered sample context can be written as

$$\widehat{\beta} = \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \boldsymbol{X}_g' \boldsymbol{Y}_g\right)$$

where $g = 1, ..., G$ indexes the cluster. Instead of leave-one-out estimators, it is natural to use delete-cluster estimators, which delete one cluster at a time. They take the form (4.58):

$$\widehat{\beta}_{(-g)} = \widehat{\beta} - \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g' \widetilde{\boldsymbol{e}}_g$$

where

$$\widetilde{\boldsymbol{e}}_g = \left(\boldsymbol{I}_{n_g} - \boldsymbol{X}_g \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}_g'\right)^{-1} \widehat{\boldsymbol{e}}_g$$
$$\widehat{\boldsymbol{e}}_g = \boldsymbol{Y}_g - \boldsymbol{X}_g \widehat{\beta}.$$

The delete-cluster jackknife estimator of the variance of $\widehat{\beta}$ is

$$\widehat{V}_{\widehat{\beta}}^{\text{jack}} = \frac{G-1}{G}\sum_{g=1}^{G}\left(\widehat{\beta}_{(-g)} - \overline{\beta}\right)\left(\widehat{\beta}_{(-g)} - \overline{\beta}\right)'$$
$$\overline{\beta} = \frac{1}{G}\sum_{g=1}^{G}\widehat{\beta}_{(-g)}.$$

We call $\widehat{V}_{\widehat{\beta}}^{\text{jack}}$ a **cluster-robust jackknife estimator of variance**.

Using the same approximations as the previous section we can show that the delete-cluster jackknife estimator is asymptotically equivalent to the cluster-robust covariance matrix estimator (4.59) calculated with the delete-cluster prediction errors. This verifies that the delete-cluster jackknife is the appropriate jackknife approach for clustered dependence.

For parameters which are functions $\widehat{\theta} = r(\widehat{\beta})$ of the least squares estimator, the delete-cluster jackknife estimator of the variance of $\widehat{\theta}$ is

$$\widehat{V}_{\widehat{\theta}}^{\text{jack}} = \frac{G-1}{G}\sum_{g=1}^{G}\left(\widehat{\theta}_{(-g)} - \overline{\theta}\right)\left(\widehat{\theta}_{(-g)} - \overline{\theta}\right)'$$
$$\widehat{\theta}_{(-i)} = r(\widehat{\beta}_{(-g)})$$
$$\overline{\theta} = \frac{1}{G}\sum_{g=1}^{G}\widehat{\theta}_{(-g)}.$$

Using a mean-value expansion we can show that this estimator is asymptotically equivalent to the delta-method cluster-robust covariance matrix estimator for $\widehat{\theta}$. This shows that the jackknife estimator is appropriate for covariance matrix estimation.

As in the context of i.i.d. samples, one advantage of the jackknife covariance matrix estimators is that they do not require the user to make a technical calculation of the asymptotic distribution. A downside is an increase in computation cost, as G separate regressions are effectively estimated.

In Stata, jackknife standard errors for coefficient estimates with clustered observations are obtained by using the options `cluster(id) vce(jackknife)` where `id` denotes the cluster variable.

10.6 The Bootstrap Algorithm

The bootstrap is a powerful approach to inference and is due to the pioneering work of Efron (1979). There are many textbook and monograph treatments of the bootstrap, including Efron (1982), Hall (1992), Efron and Tibshirani (1993), Shao and Tu (1995), and Davison and Hinkley (1997). Reviews for econometricians are provided by Hall (1994) and Horowitz (2001)

There are several ways to describe or define the bootstrap and there are several forms of the bootstrap. We start in this section by describing the basic nonparametric bootstrap algorithm. In subsequent sections we give more formal definitions of the bootstrap as well as theoretical justifications.

Briefly, the bootstrap distribution is obtained by estimation on independent samples created by i.i.d. sampling (sampling with replacement) from the original dataset.

To understand this it is useful to start with the concept of sampling with replacement from the dataset. To continue the empirical example used earlier in the chapter we focus on the dataset displayed in Table 3.1, which has $n = 20$ observations. Sampling from this distribution means randomly selecting one row from this table. Mathematically this is the same as randomly selecting an integer from the set $\{1, 2, ..., 20\}$. To illustrate, MATLAB has a random integer generator (the function `randi`). Using

the random number seed of 13 (an arbitrary choice) we obtain the random draw 16. This means that we draw observation number 16 from Table 3.1. Examining the table we can see that this is an individual with wage $18.75 and education of 16 years. We repeat by drawing another random integer on the set $\{1,2,...,20\}$ and this time obtain 5. This means we take observation 5 from Table 3.1, which is an individual with wage $33.17 and education of 16 years. We continue until we have $n = 20$ such draws. This random set of observations are {16, 5, 17, 20, 20, 10, 13, 16, 13, 15, 1, 6, 2, 18, 8, 14, 6, 7, 1, 8}. We call this the **bootstrap sample**.

Notice that the observations 1, 6, 8, 13, 16, 20 each appear twice in the bootstrap sample, and the observations 3, 4, 9, 11, 12, 19 do not appear at all. That is okay. In fact, it is necessary for the bootstrap to work. This is because we are **drawing with replacement**. (If we instead made draws without replacement then the constructed dataset would have exactly the same observations as in Table 3.1, only in different order.) We can also ask the question "What is the probability that an individual observation will appear at least once in the bootstrap sample?" The answer is

$$\begin{aligned}\mathbb{P}\left[\text{Observation in Bootstrap Sample}\right] &= 1-\left(1-\frac{1}{n}\right)^n \\ &\to 1-e^{-1} \simeq 0.632.\end{aligned} \tag{10.6}$$

The limit holds as $n \to \infty$. The approximation 0.632 is excellent even for small n. For example, when $n = 20$ the probability (10.6) is 0.641. These calculations show that an individual observation is in the bootstrap sample with probability near 2/3.

Once again, the bootstrap sample is the constructed dataset with the 20 observations drawn randomly from the original sample. Notationally, we write the i^{th} bootstrap observation as $\left(Y_i^*, X_i^*\right)$ and the bootstrap sample as $\{\left(Y_1^*, X_1^*\right), ..., \left(Y_n^*, X_n^*\right)\}$. In our present example with Y denoting the log wage the bootstrap sample is

$$\{\left(Y_1^*, X_1^*\right), ..., \left(Y_n^*, X_n^*\right)\} = \{(2.93, 16), (3.50, 16) ..., (3.76, 18)\}.$$

The bootstrap estimate $\widehat{\beta}^*$ is obtained by applying the least squares estimation formula to the bootstrap sample. Thus we regress Y^* on X^*. The other bootstrap estimates, in our example $\widehat{\sigma}^{2*}$ and $\widehat{\mu}^*$, are obtained by applying their estimation formulae to the bootstrap sample as well. Writing $\widehat{\theta}^* = \left(\widehat{\beta}_1^*, \widehat{\beta}_2^*, \widehat{\sigma}^{*2}, \widehat{\mu}^*\right)'$ we have the bootstrap estimate of the parameter vector $\theta = \left(\beta_1, \beta_2, \sigma^2, \mu\right)'$. In our example (the bootstrap sample described above) $\widehat{\theta}^* = (0.195, 0.113, 0.107, 26.7)'$. This is one draw from the bootstrap distribution of the estimates.

The estimate $\widehat{\theta}^*$ as described is one random draw from the distribution of estimates obtained by i.i.d. sampling from the original data. With one draw we can say relatively little. But we can repeat this exercise to obtain multiple draws from this bootstrap distribution. To distinguish between these draws we index the bootstrap samples by $b = 1, ..., B$, and write the bootstrap estimates as $\widehat{\theta}_b^*$ or $\widehat{\theta}^*(b)$.

To continue our illustration we draw 20 more random integers {19, 5, 7, 19, 1, 2, 13, 18, 1, 15, 17, 2, 14, 11, 10, 20, 1, 5, 15, 7} and construct a second bootstrap sample. On this sample we again estimate the parameters and obtain $\widehat{\theta}^*(2) = (0.175, 0.52, 0.124, 29.3)'$. This is a second random draw from the distribution of $\widehat{\theta}^*$. We repeat this B times, storing the parameter estimates $\widehat{\theta}^*(b)$. We have thus created a new dataset of bootstrap draws $\left\{\widehat{\theta}^*(b) : b = 1, ..., B\right\}$. By construction the draws are independent across b and identically distributed.

The number of bootstrap draws, B, is often called the "number of bootstrap replications". Typical choices for B are 1000, 5000, and 10,000. We discuss selecting B later, but roughly speaking, larger B results in a more precise estimate at an increased computation cost. For our application we set $B =$ 10,000.

To illustrate, Figure 13.1 displays the densities of the distributions of the bootstrap estimates $\widehat{\beta}_1^*$ and $\widehat{\mu}^*$ across 10,000 draws. The dashed lines show the point estimate. You can notice that the density for $\widehat{\beta}_1^*$ is slightly skewed to the left.

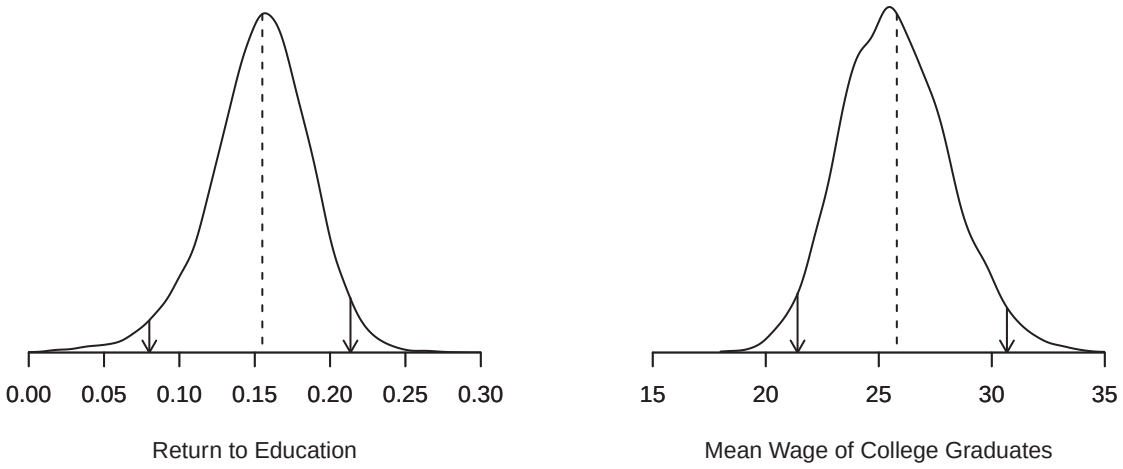


Figure 10.1: Bootstrap Distributions of $\widehat{\beta}_1^*$ and $\widehat{\mu}^*$

10.7 Bootstrap Variance and Standard Errors

Given the bootstrap draws we can estimate features of the bootstrap distribution. The **bootstrap estimator of variance** of an estimator $\widehat{\theta}$ is the sample variance across the bootstrap draws $\widehat{\theta}^*(b)$. It equals

$$\widehat{V}_{\widehat{\theta}}^{\text{boot}} = \frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\theta}^*(b) - \overline{\theta}^*\right)\left(\widehat{\theta}^*(b) - \overline{\theta}^*\right)' \tag{10.7}$$

$$\overline{\theta}^* = \frac{1}{B}\sum_{b=1}^{B}\widehat{\theta}^*(b).$$

For a scalar estimator $\widehat{\theta}$ the **bootstrap standard error** is the square root of the bootstrap estimator of variance:

$$s_{\widehat{\theta}}^{\text{boot}} = \sqrt{\widehat{V}_{\widehat{\theta}}^{\text{boot}}}.$$

This is a very simple statistic to calculate and is the most common use of the bootstrap in applied econometric practice. A caveat (discussed in more detail in Section 10.15) is that in many cases it is better to use a trimmed estimator.

Standard errors are conventionally reported to convey the precision of the estimator. They are also commonly used to construct confidence intervals. Bootstrap standard errors can be used for this purpose. The **normal-approximation bootstrap confidence interval** is

$$C^{\text{nb}} = \left[\widehat{\theta} - z_{1-\alpha/2}s_{\widehat{\theta}}^{\text{boot}},\quad \widehat{\theta} + z_{1-\alpha/2}s_{\widehat{\theta}}^{\text{boot}}\right]$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the $\mathrm{N}(0,1)$ distribution. This interval C^{nb} is identical in format to an asymptotic confidence interval, but with the bootstrap standard error replacing the asymptotic standard error. C^{nb} is the default confidence interval reported by Stata when the bootstrap has been used to calculate standard errors. However, the normal-approximation interval is in general a poor choice for confidence interval construction as it relies on the normal approximation to the t-ratio which can be inaccurate in finite samples. There are other methods – such as the bias-corrected percentile method to be discussed in Section 10.17 – which are just as simple to compute but have better performance. In general, bootstrap standard errors should be used as estimates of precision rather than as tools to construct confidence intervals.

Since B is finite, all bootstrap statistics, such as $\widehat{V}_{\widehat{\theta}}^{\mathrm{boot}}$, are estimates and hence random. Their values will vary across different choices for B and simulation runs (depending on how the simulation seed is set). Thus you should not expect to obtain the exact same bootstrap standard errors as other researchers when replicating their results. They should be similar (up to simulation sampling error) but not precisely the same.

In Table 10.2 we report the four parameter estimates introduced in Section 10.2 along with asymptotic, jackknife and bootstrap standard errors. We also report four bootstrap confidence intervals which will be introduced in subsequent sections.

For these four estimators we can see that the bootstrap standard errors are quite similar to the asymptotic and jackknife standard errors. The most noticable difference arises for $\widehat{\beta}_2$, where the bootstrap standard error is about 10% larger than the asymptotic standard error.

Table 10.2: Comparison of Methods

	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\sigma}^2$	$\widehat{\mu}$
Estimate	0.155	0.698	0.144	25.80
Asymptotic s.e.	(0.031)	(0.493)	(0.043)	(2.29)
Jackknife s.e.	(0.032)	(0.514)	(0.046)	(2.39)
Bootstrap s.e.	(0.034)	(0.548)	(0.041)	(2.38)
95% Percentile Interval	[0.08, 0.21]	[−0.27, 1.91]	[0.06, 0.22]	[21.4, 30.7]
95% BC Percentile Interval	[0.08, 0.21]	[−0.25, 1.93]	[0.09, 0.28]	[22.0, 31.5]
95% BC_a Percentile Interval	[0.08, 0.21]	[−0.25, 1.93]	[0.09, 0.28]	[22.0, 31.5]
95% Percentile-t Interval	[0.09, 0.21]	[−0.20, 1.81]	[0.08, 0.34]	[21.6, 32.2]

In Stata, bootstrap standard errors for coefficient estimates in many models are obtained by the `vce(bootstrap, reps(#))` option, where `#` is the number of bootstrap replications. For nonlinear functions of the coefficients or other estimators the `bootstrap` command can be combined with any other command to obtain bootstrap standard errors. Synonyms for `bootstrap` are `bstrap` and `bs`.

To illustrate, below we list the Stata commands which will calculate[1] the bootstrap standard errors listed above.

[1] They will not *precisely* replicate the standard errors since those in Table 10.2 were produced in Matlab which uses a different random number sequence.

Stata Commands

```
reg wage education if mbf12 == 1, vce(bootstrap, reps(10000))
bs (e(rss)/e(N)), reps(10000): reg wage education if mbf12 == 1
bs (exp(16*_b[education]+_b[_cons]+e(rss)/e(N)/2)), reps(10000): ///
    reg wage education if mbf12 == 1
```

10.8 Percentile Interval

The second most common use of bootstrap methods is for confidence intervals. There are multiple bootstrap methods to form confidence intervals. A popular and simple method is called the **percentile interval**. It is based on the quantiles of the bootstrap distribution.

In Section 10.6 we described the bootstrap algorithm which creates an i.i.d. sample of bootstrap estimates $\{\widehat{\theta}_1^*, \widehat{\theta}_2^*, ..., \widehat{\theta}_B^*\}$ corresponding to an estimator $\widehat{\theta}$ of a parameter θ. We focus on the case of a scalar parameter θ.

For any $0 < \alpha < 1$ we can calculate the empirical quantile q_α^* of these bootstrap estimates. This is the number such that $n\alpha$ bootstrap estimates are smaller than q_α^*, and is typically calculated by taking the $n\alpha^{th}$ order statistic of the $\widehat{\theta}_b^*$. See Section 11.13 of *Probability and Statistics for Economists* for a precise discussion of empirical quantiles and common quantile estimators.

The percentile bootstrap $100(1-\alpha)\%$ confidence interval is

$$C^{\text{pc}} = \left[q_{\alpha/2}^*, q_{1-\alpha/2}^*\right]. \tag{10.8}$$

For example, if $B = 1000$, $\alpha = 0.05$, and the empirical quantile estimator is used, then $C^{\text{pc}} = \left[\widehat{\theta}_{(25)}^*, \widehat{\theta}_{(975)}^*\right]$.

To illustrate, the 0.025 and 0.975 quantiles of the bootstrap distributions of $\widehat{\beta}_1^*$ and $\widehat{\mu}^*$ are indicated in Figure 13.1 by the arrows. The intervals between the arrows are the 95% percentile intervals.

The percentile interval has the convenience that it does not require calculation of a standard error. This is particularly convenient in contexts where asymptotic standard error calculation is complicated, burdensome, or unknown. C^{pc} is a simple by-product of the bootstrap algorithm and does not require meaningful computational cost above that required to calculate the bootstrap standard error.

The percentile interval has the useful property that it is **transformation-respecting**. Take a monotone parameter transformation $m(\theta)$. The percentile interval for $m(\theta)$ is simply the percentile interval for θ mapped by $m(\theta)$. That is, if $\left[q_{\alpha/2}^*, q_{1-\alpha/2}^*\right]$ is the percentile interval for θ, then $\left[m\left(q_{\alpha/2}^*\right), m\left(q_{1-\alpha/2}^*\right)\right]$ is the percentile interval for $m(\theta)$. This property follows directly from the equivariance property of sample quantiles. Many confidence-interval methods, such as the delta-method asymptotic interval and the normal-approximation interval C^{nb}, do not share this property.

To illustrate the usefulness of the transformation-respecting property consider the variance σ^2. In some cases it is useful to report the variance σ^2 and in other cases it is useful to report the standard deviation σ. Thus we may be interested in confidence intervals for σ^2 or σ. To illustrate, the asymptotic 95% normal confidence interval for σ^2 which we calculate from Table 13.2 is $[0.060, 0.228]$. Taking square roots we obtain an interval for σ of $[0.244, 0.477]$. Alternatively, the delta method standard error for $\widehat{\sigma} = 0.379$ is 0.057, leading to an asymptotic 95% confidence interval for σ of $[0.265, 0.493]$ which is different. This shows that the delta method is not transformation-respecting. In contrast, the 95% percentile interval for σ^2 is [0.062, 0.220] and that for σ is [0.249, 0.469] which is identical to the square roots of the interval for σ^2.

The bootstrap percentile intervals for the four estimators are reported in Table 13.2.

In Stata, percentile confidence intervals can be obtained by using the command `estat bootstrap, percentile` or the command `estat bootstrap, all` after an estimation command which calculates standard errors via the bootstrap.

10.9 The Bootstrap Distribution

For applications it is often sufficient if one understands the bootstrap as an algorithm. However, for theory it is more useful to view the bootstrap as a specific estimator of the sampling distribution. For this it is useful to introduce some additional notation.

The key is that the distribution of any estimator or statistic is determined by the distribution of the data. While the latter is unknown it can be estimated by the empirical distribution of the data. This is what the bootstrap does.

To fix notation, let F denote the distribution of an individual observation W. (In regression, W is the pair (Y, X).) Let $G_n(u, F)$ denote the distribution of an estimator $\widehat{\theta}$. That is,

$$G_n(u, F) = \mathbb{P}\left[\widehat{\theta} \leq u \mid F\right].$$

We write the distribution G_n as a function of n and F since the latter (generally) affect the distribution of $\widehat{\theta}$. We are interested in the distribution G_n. For example, we want to know its variance to calculate a standard error or its quantiles to calculate a percentile interval.

In principle, if we knew the distribution F we should be able to determine the distribution G_n. In practice there are two barriers to implementation. The first barrier is that the calculation of $G_n(u, F)$ is generally infeasible except in certain special cases such as the normal regression model. The second barrier is that in general we do not know F.

The bootstrap simultaneously circumvents these two barriers by two clever ideas. First, the bootstrap proposes estimation of F by the empirical distribution function (EDF) F_n, which is the simplest nonparametric estimator of the joint distribution of the observations. The EDF is $F_n(w) = n^{-1}\sum_{i=1}^{n} \mathbb{1}\{W_i \leq w\}$. (See Section 11.2 of *Probability and Statistics for Economists* for details and properties.) Replacing F with F_n we obtain the idealized bootstrap estimator of the distribution of $\widehat{\theta}$

$$G_n^*(u) = G_n(u, F_n). \tag{10.9}$$

The bootstrap's second clever idea is to estimate G_n^* by simulation. This is the bootstrap algorithm described in the previous sections. The essential idea is that simulation from F_n is sampling with replacement from the original data, which is computationally simple. Applying the estimation formula for $\widehat{\theta}$ we obtain i.i.d. draws from the distribution $G_n^*(u)$. By making a large number B of such draws we can estimate any feature of G_n^* of interest. The bootstrap combines these two ideas: (1) estimate $G_n(u, F)$ by $G_n(u, F_n)$; (2) estimate $G_n(u, F_n)$ by simulation. These ideas are intertwined. Only by considering these steps together do we obtain a feasible method.

The way to think about the connection between G_n and G_n^* is as follows. G_n is the distribution of the estimator $\widehat{\theta}$ obtained when the observations are sampled i.i.d. from the population distribution F. G_n^* is the distribution of the same statistic, denoted $\widehat{\theta}^*$, obtained when the observations are sampled i.i.d. from the empirical distribution F_n. It is useful to conceptualize the "universe" which separately generates the dataset and the bootstrap sample. The "sampling universe" is the population distribution F. In this universe the true parameter is θ. The "bootstrap universe" is the empircal distribution F_n. When drawing from the bootstrap universe we are treating F_n as if it is the true distribution. Thus anything which is true about F_n should be treated as true in the bootstrap universe. In the bootstrap universe the "true" value of the parameter θ is the value determined by the EDF F_n. In most cases this is the estimate $\widehat{\theta}$. It is the true value of the coefficient when the true distribution is F_n.

We now carefully explain the connection with the bootstrap algorithm as previously described.

First, observe that sampling with replacement from the sample $\{Y_1, ..., Y_n\}$ is identical to sampling from the EDF F_n. This is because the EDF is the probability distribution which puts probability mass $1/n$ on each observation. Thus sampling from F_n means sampling an observation with probability $1/n$, which is sampling with replacement.

Second, observe that the bootstrap estimator $\widehat{\theta}^*$ described here is identical to the bootstrap algorithm described in Section 10.6. That is, $\widehat{\theta}^*$ is the random vector generated by applying the estimator formula $\widehat{\theta}$ to samples obtained by random sampling from F_n.

Third, observe that the distribution of these bootstrap estimators is the bootstrap distribution (10.9). This is a precise equality. That is, the bootstrap algorithm generates i.i.d. samples from F_n, and when the estimators are applied we obtain random variables $\widehat{\theta}^*$ with the distribution G_n^*.

Fourth, observe that the bootstrap statistics described earlier – bootstrap variance, standard error, and quantiles – are estimators of the corresponding features of the bootstrap distribution G_n^*.

This discussion is meant to carefully describe why the notation $G_n^*(u)$ is useful to help understand the properties of the bootstrap algorithm. Since F_n is the natural nonparametric estimator of the unknown distribution F, $G_n^*(u) = G_n(u, F_n)$ is the natural plug-in estimator of the unknown $G_n(u, F)$. Furthermore, because F_n is uniformly consistent for F by the Glivenko-Cantelli Lemma (Theorem 18.8 in *Probability and Statistics for Economists*) we also can expect $G_n^*(u)$ to be consistent for $G_n(u)$. Making this precise is a bit challenging since F_n and G_n are functions. In the next several sections we develop an asymptotic distribution theory for the bootstrap distribution based on extending asymptotic theory to the case of conditional distributions.

10.10 The Distribution of the Bootstrap Observations

Let Y^* be a random draw from the sample $\{Y_1, ..., Y_n\}$. What is the distribution of Y^*?

Since we are fixing the observations, the correct question is: What is the *conditional* distribution of Y^*, conditional on the observed data? The empirical distribution function F_n summarizes the information in the sample, so equivalently we are talking about the distribution conditional on F_n. Consequently we will write the bootstrap probability function and expectation as

$$\begin{aligned}\mathbb{P}^*\left[Y^* \le x\right] &= \mathbb{P}\left[Y^* \le x \mid F_n\right] \\ \mathbb{E}^*\left[Y^*\right] &= \mathbb{E}\left[Y^* \mid F_n\right].\end{aligned}$$

Notationally, the starred distribution and expectation are conditional given the data.

The (conditional) distribution of Y^* is the empirical distribution function F_n, which is a discrete distribution with mass points $1/n$ on each observation Y_i. Thus even if the original data come from a continuous distribution, the bootstrap data distribution is discrete.

The (conditional) mean and variance of Y^* are calculated from the EDF, and equal the sample mean and variance of the data. The mean is

$$\mathbb{E}^*\left[Y^*\right] = \sum_{i=1}^{n} Y_i \mathbb{P}^*\left[Y^* = Y_i\right] = \sum_{i=1}^{n} Y_i \frac{1}{n} = \overline{Y} \tag{10.10}$$

and the variance is

$$\begin{aligned}\operatorname{var}^*\left[Y^*\right] &= \mathbb{E}^*\left[Y^*Y^{*\prime}\right] - \left(\mathbb{E}^*\left[Y^*\right]\right)\left(\mathbb{E}^*\left[Y^*\right]\right)' \\ &= \sum_{i=1}^{n} Y_i Y_i' \mathbb{P}^*\left[Y^* = Y_i\right] - \overline{Y}\,\overline{Y}' \\ &= \sum_{i=1}^{n} Y_i Y_i' \frac{1}{n} - \overline{Y}\,\overline{Y}' \\ &= \widehat{\Sigma}. \end{aligned} \tag{10.11}$$

To summarize, the conditional distribution of Y^*, given F_n, is the discrete distribution on $\{Y_1, ..., Y_n\}$ with mean $\overline{Y}$ and covariance matrix $\widehat{\Sigma}$.

We can extend this analysis to any integer moment r. Assume Y is scalar. The r^{th} moment of Y^* is

$$\mu_r^{*\prime} = \mathbb{E}^*\left[Y^{*r}\right] = \sum_{i=1}^{n} Y_i^r \mathbb{P}^*\left[Y^* = Y_i\right] = \frac{1}{n}\sum_{i=1}^{n} Y_i^r = \widehat{\mu}_r',$$

the r^{th} sample moment. The r^{th} central moment of Y^* is

$$\mu_r^* = \mathbb{E}^*\left[\left(Y^* - \overline{Y}\right)^r\right] = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^r = \widehat{\mu}_r,$$

the r^{th} central sample moment. Similarly, the r^{th} cumulant of Y^* is $\kappa_r^* = \widehat{\kappa}_r$, the r^{th} sample cumulant.

10.11 The Distribution of the Bootstrap Sample Mean

The bootstrap sample mean is

$$\overline{Y}^* = \frac{1}{n}\sum_{i=1}^{n} Y_i^*.$$

We can calculate its (conditional) mean and variance. The mean is

$$\mathbb{E}^*\left[\overline{Y}^*\right] = \mathbb{E}^*\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^*\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}^*\left[Y_i^*\right] = \frac{1}{n}\sum_{i=1}^{n}\overline{Y} = \overline{Y}. \tag{10.12}$$

using (10.10). Thus the bootstrap sample mean $\overline{Y}^*$ has a distribution centered at the sample mean $\overline{Y}$. This is because the bootstrap observations Y_i^* are drawn from the bootstrap universe, which treats the EDF as the truth, and the mean of the latter distribution is $\overline{Y}$.

The (conditional) variance of the bootstrap sample mean is

$$\operatorname{var}^*\left[\overline{Y}^*\right] = \operatorname{var}^*\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^*\right] = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{var}^*\left[Y_i^*\right] = \frac{1}{n^2}\sum_{i=1}^{n}\widehat{\Sigma} = \frac{1}{n}\widehat{\Sigma} \tag{10.13}$$

using (10.11). In the scalar case, $\operatorname{var}^*\left[\overline{Y}^*\right] = \widehat{\sigma}^2/n$. This shows that the bootstrap variance of $\overline{Y}^*$ is precisely described by the sample variance of the original observations. Again, this is because the bootstrap observations Y_i^* are drawn from the bootstrap universe.

We can extend this to any integer moment r. Assume Y is scalar. Define the normalized bootstrap sample mean $Z_n^* = \sqrt{n}\left(\overline{Y}^* - \overline{Y}\right)$. Using expressions from Section 6.17 of *Probability and Statistics for*

Economists, the 3^{rd} through 6^{th} conditional moments of Z_n^* are

$$\begin{aligned}
\mathbb{E}^*\left[Z_n^{*3}\right] &= \widehat{\kappa}_3/n^{1/2} \\
\mathbb{E}^*\left[Z_n^{*4}\right] &= \widehat{\kappa}_4/n + 3\widehat{\kappa}_2^2 \\
\mathbb{E}^*\left[Z_n^{*5}\right] &= \widehat{\kappa}_5/n^{3/2} + 10\widehat{\kappa}_3\widehat{\kappa}_2/n^{1/2} \\
\mathbb{E}^*\left[Z_n^{*6}\right] &= \widehat{\kappa}_6/n^2 + \left(15\widehat{\kappa}_4\kappa_2 + 10\widehat{\kappa}_3^2\right)/n + 15\widehat{\kappa}_2^3
\end{aligned} \tag{10.14}$$

where $\widehat{\kappa}_r$ is the r^{th} sample cumulant. Similar expressions can be derived for higher moments. The moments (10.14) are exact, not approximations.

10.12 Bootstrap Asymptotics

The bootstrap mean $\overline{Y}^*$ is a sample average over n i.i.d. random variables, so we might expect it to converge in probability to its expectation. Indeed, this is the case, but we have to be a bit careful since the bootstrap mean has a conditional distribution (given the data) so we need to define convergence in probability for conditional distributions.

Definition 10.1 We say that a random vector Z_n^* **converges in bootstrap probability** to Z as $n \to \infty$, denoted $Z_n^* \underset{p^*}{\longrightarrow} Z$, if for all $\epsilon > 0$

$$\mathbb{P}^*\left[\left\|Z_n^* - Z\right\| > \epsilon\right] \underset{p}{\longrightarrow} 0.$$

To understand this definition recall that conventional convergence in probability $Z_n \underset{p}{\longrightarrow} Z$ means that for a sufficiently large sample size n, the probability is high that Z_n is arbitrarily close to its limit Z. In contrast, Definition 10.1 says $Z_n^* \underset{p^*}{\longrightarrow} Z$ means that for a sufficiently large n, the probability is high that the conditional probability that Z_n^* is close to its limit Z is high. Note that there are two uses of probability – both unconditional and conditional.

Our label "convergence in bootstrap probability" is a bit unusual. The label used in much of the statistical literature is "convergence in probability, in probability" but that seems like a mouthful. That literature more often focuses on the related concept of "convergence in probability, almost surely" which holds if we replace the "$\underset{p}{\longrightarrow}$" convergence with almost sure convergence. We do not use this concept in this chapter as it is an unnecessary complication.

While we have stated Definition 10.1 for the specific conditional probability distribution $\mathbb{P}^*$, the idea is more general and can be used for any conditional distribution and any sequence of random vectors.

The following may seem obvious but it is useful to state for clarity. Its proof is given in Section 10.31.

Theorem 10.1 If $Z_n \underset{p}{\longrightarrow} Z$ as $n \to \infty$ then $Z_n \underset{p^*}{\longrightarrow} Z$.

Given Definition 10.1, we can establish a law of large numbers for the bootstrap sample mean.

Theorem 10.2 Bootstrap WLLN. If Y_i are independent and uniformly integrable then $\overline{Y}^* - \overline{Y} \underset{p^*}{\longrightarrow} 0$ and $\overline{Y}^* \underset{p^*}{\longrightarrow} \mu = \mathbb{E}[Y]$ as $n \to \infty$.

The proof (presented in Section 10.31) is somewhat different from the classical case as it is based on the Marcinkiewicz WLLN (Theorem 10.20, presented in Section 10.31).

Notice that the conditions for the bootstrap WLLN are the same for the conventional WLLN. Notice as well that we state two related but slightly different results. The first is that the difference between the bootstrap sample mean $\overline{Y}^*$ and the sample mean $\overline{Y}$ diminishes as the sample size diverges. The second result is that the bootstrap sample mean converges to the population mean μ. The latter is not surprising (since the sample mean $\overline{Y}$ converges in probability to μ) but it is constructive to be precise since we are dealing with a new convergence concept.

Theorem 10.3 Bootstrap Continuous Mapping Theorem. If $Z_n^* \underset{p^*}{\longrightarrow} c$ as $n \to \infty$ and $g(\cdot)$ is continuous at c, then $g(Z_n^*) \underset{p^*}{\longrightarrow} g(c)$ as $n \to \infty$.

The proof is essentially identical to that of Theorem 6.6 so is omitted.

We next would like to show that the bootstrap sample mean is asymptotically normally distributed, but for that we need a definition of convergence for conditional distributions.

Definition 10.2 Let Z_n^* be a sequence of random vectors with conditional distributions $G_n^*(x) = \mathbb{P}^*\left[Z_n^* \le x\right]$. We say that Z_n^* **converges in bootstrap distribution** to Z as $n \to \infty$, denoted $Z_n^* \underset{d^*}{\longrightarrow} Z$, if for all x at which $G(x) = \mathbb{P}[Z \le x]$ is continuous, $G_n^*(x) \underset{p}{\longrightarrow} G(x)$ as $n \to \infty$.

The difference with the conventional definition is that Definition 10.2 treats the conditional distribution as random. An alternative label for Definition 10.2 is "convergence in distribution, in probability".

We now state a CLT for the bootstrap sample mean, with a proof given in Section 10.31.

Theorem 10.4 Bootstrap CLT. If Y_i are i.i.d., $\mathbb{E}\|Y\|^2 < \infty$, and $\Sigma = \operatorname{var}[Y] > 0$, then as $n \to \infty$, $\sqrt{n}\left(\overline{Y}^* - \overline{Y}\right) \underset{d^*}{\longrightarrow} \mathrm{N}(0, \Sigma)$.

Theorem 10.4 shows that the normalized bootstrap sample mean has the same asymptotic distribution as the sample mean. Thus the bootstrap distribution is asymptotically the same as the sampling distribution. A notable difference, however, is that the bootstrap sample mean is normalized by centering at the sample mean, not at the population mean. This is because $\overline{Y}$ is the true mean in the bootstrap universe.

We next state the distributional form of the continuous mapping theorem for bootstrap distributions and the Bootstrap Delta Method.

Theorem 10.5 Bootstrap Continuous Mapping Theorem
If $Z_n^* \underset{d^*}{\longrightarrow} Z$ as $n \to \infty$ and $g : \mathbb{R}^m \to \mathbb{R}^k$ has the set of discontinuity points D_g such that $\mathbb{P}^*\left[Z^* \in D_g\right] = 0$, then $g(Z_n^*) \underset{d^*}{\longrightarrow} g(Z)$ as $n \to \infty$.

Theorem 10.6 Bootstrap Delta Method:
If $\widehat{\mu} \underset{p}{\longrightarrow} \mu$, $\sqrt{n}\left(\widehat{\mu}^* - \widehat{\mu}\right) \underset{d^*}{\longrightarrow} \xi$, and $g(u)$ is continuously differentiable in a neighborhood of μ, then as $n \to \infty$

$$\sqrt{n}\left(g\left(\widehat{\mu}^*\right) - g(\widehat{\mu})\right) \underset{d^*}{\longrightarrow} \boldsymbol{G}'\xi$$

where $\boldsymbol{G}(x) = \frac{\partial}{\partial x} g(x)'$ and $\boldsymbol{G} = \boldsymbol{G}(\mu)$. In particular, if $\xi \sim \mathrm{N}(0, \boldsymbol{V})$ then as $n \to \infty$

$$\sqrt{n}\left(g\left(\widehat{\mu}^*\right) - g(\widehat{\mu})\right) \underset{d^*}{\longrightarrow} \mathrm{N}\left(0, \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}\right).$$

For a proof, see Exercise 10.7.

We state an analog of Theorem 6.10, which presented the asymptotic distribution for general smooth functions of sample means, which covers most econometric estimators.

Theorem 10.7 Under the assumptions of Theorem 6.10, that is, if Y_i is i.i.d., $\mu = \mathbb{E}[h(Y)]$, $\theta = g\left(\mu\right)$, $\mathbb{E}\|h(Y)\|^2 < \infty$, and $\boldsymbol{G}(x) = \dfrac{\partial}{\partial x} g(x)'$ is continuous in a neighborhood of μ, for $\widehat{\theta} = g\left(\widehat{\mu}\right)$ with $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n h(Y_i)$ and $\widehat{\theta}^* = g\left(\widehat{\mu}^*\right)$ with $\widehat{\mu}^* = \frac{1}{n}\sum_{i=1}^n h\left(Y_i^*\right)$, as $n \to \infty$

$$\sqrt{n}\left(\widehat{\theta}^* - \widehat{\theta}\right) \underset{d^*}{\longrightarrow} \mathrm{N}(0, \boldsymbol{V}_\theta)$$

where $\boldsymbol{V}_\theta = \boldsymbol{G}'\boldsymbol{V}\boldsymbol{G}$, $\boldsymbol{V} = \mathbb{E}\left[\left(h(Y) - \mu\right)\left(h(Y) - \mu\right)'\right]$ and $\boldsymbol{G} = \boldsymbol{G}\left(\mu\right)$.

For a proof, see Exercise 10.8.

Theorem 10.7 shows that the asymptotic distribution of the bootstrap estimator $\widehat{\theta}^*$ is identical to that of the sample estimator $\widehat{\theta}$. This means that we can learn the distribution of $\widehat{\theta}$ from the bootstrap distribution, and hence perform asymptotically correct inference.

For some bootstrap applications we use bootstrap estimates of variance. The plug-in estimator of $\boldsymbol{V}_\theta$ is $\widehat{\boldsymbol{V}}_\theta = \widehat{\boldsymbol{G}}'\widehat{\boldsymbol{V}}\widehat{\boldsymbol{G}}$ where $\widehat{\boldsymbol{G}} = \boldsymbol{G}\left(\widehat{\mu}\right)$ and

$$\widehat{\boldsymbol{V}} = \frac{1}{n}\sum_{i=1}^n \left(h(Y_i) - \widehat{\mu}\right)\left(h(Y_i) - \widehat{\mu}\right)'.$$

The bootstrap version is

$$\begin{aligned}\widehat{\boldsymbol{V}}_\theta^* &= \widehat{\boldsymbol{G}}^{*\prime}\widehat{\boldsymbol{V}}^*\widehat{\boldsymbol{G}}^* \\ \widehat{\boldsymbol{G}}^* &= \boldsymbol{G}\left(\widehat{\mu}^*\right) \\ \widehat{\boldsymbol{V}}^* &= \frac{1}{n}\sum_{i=1}^n \left(h\left(Y_i^*\right)-\widehat{\mu}^*\right)\left(h\left(Y_i^*\right)-\widehat{\mu}^*\right)'.\end{aligned}$$

Application of the bootstrap WLLN and bootstrap CMT show that $\widehat{\boldsymbol{V}}_\theta^*$ is consistent for $\boldsymbol{V}_\theta$.

Theorem 10.8 Under the assumptions of Theorem 10.7, $\widehat{\boldsymbol{V}}_\theta^* \underset{p^*}{\longrightarrow} \boldsymbol{V}_\theta$ as $n \to \infty$.

For a proof, see Exercise 10.9.

10.13 Consistency of the Bootstrap Estimate of Variance

Recall the definition (10.7) of the bootstrap estimator of variance $\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{\text{boot}}$ of an estimator $\widehat{\theta}$. In this section we explore conditions under which $\widehat{\boldsymbol{V}}_{\widehat{\theta}}^{\text{boot}}$ is consistent for the asymptotic variance of $\widehat{\theta}$.

To do so it is useful to focus on a normalized version of the estimator so that the asymptotic variance is not degenerate. Suppose that for some sequence a_n we have

$$Z_n = a_n\left(\widehat{\theta}-\theta\right) \underset{d}{\longrightarrow} \xi \tag{10.15}$$

and

$$Z_n^* = a_n\left(\widehat{\theta}^*-\widehat{\theta}\right) \underset{d^*}{\longrightarrow} \xi \tag{10.16}$$

for some limit distribution ξ. That is, for some normalization, both $\widehat{\theta}$ and $\widehat{\theta}^*$ have the same asymptotic distribution. This is quite general as it includes the smooth function model. The conventional bootstrap estimator of the variance of Z_n is the sample variance of the bootstrap draws $\left\{Z_n^*(b) : b = 1, ..., B\right\}$. This equals the estimator (10.7) multiplied by a_n^2. Thus it is equivalent (up to scale) whether we discuss estimating the variance of $\widehat{\theta}$ or Z_n.

The bootstrap estimator of variance of Z_n is

$$\begin{aligned}\widehat{\boldsymbol{V}}_\theta^{\text{boot,B}} &= \frac{1}{B-1}\sum_{b=1}^B \left(Z_n^*(b)-\overline{Z}_n^*\right)\left(Z_n^*(b)-\overline{Z}_n^*\right)' \\ \overline{Z}_n^* &= \frac{1}{B}\sum_{b=1}^B Z_n^*(b).\end{aligned}$$

Notice that we index the estimator by the number of bootstrap replications B.

Since Z_n^* converges in bootstrap distribution to the same asymptotic distribution as Z_n, it seems reasonable to guess that the variance of Z_n^* will converge to that of ξ. However, convergence in distribution is not sufficient for convergence in moments. For the variance to converge it is also necessary for the sequence Z_n^* to be uniformly square integrable.

Theorem 10.9 If (10.15) and (10.16) hold for some sequence a_n and $\left\|Z_n^*\right\|^2$ is uniformly integrable, then as $B \to \infty$

$$\widehat{\boldsymbol{V}}_\theta^{\text{boot,B}} \xrightarrow[p^*]{} \widehat{\boldsymbol{V}}_\theta^{\text{boot}} = \text{var}\left[Z_n^*\right],$$

and as $n \to \infty$

$$\widehat{\boldsymbol{V}}_\theta^{\text{boot}} \xrightarrow[p^*]{} \boldsymbol{V}_\theta = \text{var}\left[\xi\right].$$

This raises the question: Is the normalized sequence Z_n uniformly integrable? We spend the remainder of this section exploring this question and turn in the next section to trimmed variance estimators which do not require uniform integrability.

This condition is reasonably straightforward to verify for the case of a scalar sample mean with a finite variance. That is, suppose $Z_n^* = \sqrt{n}\left(\overline{Y}^* - \overline{Y}\right)$ and $\mathbb{E}\left[Y^2\right] < \infty$. In (10.14) we calculated the exact fourth central moment of Z_n^*:

$$\mathbb{E}^*\left[Z_n^{*4}\right] = \frac{\widehat{\kappa}_4}{n} + 3\widehat{\sigma}^4 = \frac{\widehat{\mu}_4 - 3\widehat{\sigma}^4}{n} + 3\widehat{\sigma}^4$$

where $\widehat{\sigma}^2 = n^{-1}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^2$ and $\widehat{\mu}_4 = n^{-1}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^4$. The assumption $\mathbb{E}\left[Y^2\right] < \infty$ implies that $\mathbb{E}\left[\widehat{\sigma}^2\right] = O(1)$ so $\widehat{\sigma}^2 = O_p(1)$. Furthermore, $n^{-1}\widehat{\mu}_4 = n^{-2}\sum_{i=1}^n \left(Y_i - \overline{Y}\right)^4 = o_p(1)$ by the Marcinkiewicz WLLN (Theorem 10.20). It follows that

$$\mathbb{E}^*\left[Z_n^{*4}\right] = n^2\mathbb{E}^*\left[\left(\overline{Y}^* - \overline{Y}\right)^4\right] = O_p(1). \tag{10.17}$$

Theorem 6.13 shows that this implies that Z_n^{*2} is uniformly integrable. Thus if Y has a finite variance the normalized bootstrap sample mean is uniformly square integrable and the bootstrap estimate of variance is consistent by Theorem 10.9.

Now consider the smooth function model of Theorem 10.7. We can establish the following result.

Theorem 10.10 In the smooth function model of Theorem 10.7, if for some $p \geq 1$ the p^{th}-order derivatives of $g(x)$ are bounded, then $Z_n^* = \sqrt{n}\left(\widehat{\theta}^* - \widehat{\theta}\right)$ is uniformly square integrable and the bootstrap estimator of variance is consistent as in Theorem 10.9.

For a proof see Section 10.31.

This shows that the bootstrap estimate of variance is consistent for a reasonably broad class of estimators. The class of functions $g(x)$ covered by this result includes all p^{th}-order polynomials.

10.14 Trimmed Estimator of Bootstrap Variance

Theorem 10.10 showed that the bootstrap estimator of variance is consistent for smooth functions with a bounded p^{th} order derivative. This is a fairly broad class but excludes many important applications. An example is $\theta = \mu_1/\mu_2$ where $\mu_1 = \mathbb{E}[Y_1]$ and $\mu_2 = \mathbb{E}[Y_2]$. This function does not have a bounded derivative (unless μ_2 is bounded away from zero) so is not covered by Theorem 10.10.

This is more than a technical issue. When (Y_1, Y_2) are jointly normally distributed then it is known that $\widehat{\theta} = \overline{Y}_1 / \overline{Y}_2$ does not possess a finite variance. Consequently we cannot expect the bootstrap estimator of variance to perform well. (It is attempting to estimate the variance of $\widehat{\theta}$, which is infinity.)

In these cases it is preferred to use a trimmed estimator of bootstrap variance. Let $\tau_n \to \infty$ be a sequence of positive trimming numbers satisfying $\tau_n = O\left(e^{n/8}\right)$. Define the trimmed statistic

$$Z_n^{**} = Z_n^* \mathbb{1}\left\{\left\|Z_n^*\right\| \leq \tau_n\right\}.$$

The trimmed bootstrap estimator of variance is

$$\widehat{V}_\theta^{\text{boot,B},\tau} = \frac{1}{B-1}\sum_{b=1}^{B}\left(Z_n^{**}(b) - Z_n^{**}\right)\left(Z_n^{**}(b) - Z_n^{**}\right)'$$

$$Z_n^{**} = \frac{1}{B}\sum_{b=1}^{B} Z_n^{**}(b).$$

We first examine the behavior of $\widehat{V}_\theta^{\text{boot,B}}$ as the number of bootstrap replications B grows to infinity. It is a sample variance of independent bounded random vectors. Thus by the bootstrap WLLN (Theorem 10.2) $\widehat{V}_\theta^{\text{boot,B},\tau}$ converges in bootstrap probability to the variance of Z_n^{**}.

Theorem 10.11 As $B \to \infty$, $\widehat{V}_\theta^{\text{boot,B},\tau} \underset{p^*}{\longrightarrow} \widehat{V}_\theta^{\text{boot},\tau} = \text{var}\left[Z_n^{**}\right]$.

We next examine the behavior of the bootstrap estimator $\widehat{V}_\theta^{\text{boot},\tau}$ as n grows to infinity. We focus on the smooth function model of Theorem 10.7, which showed that $Z_n^* = \sqrt{n}\left(\widehat{\theta}^* - \widehat{\theta}\right) \underset{d^*}{\longrightarrow} Z \sim \text{N}(0, \boldsymbol{V}_\theta)$. Since the trimming is asymptotically negligible, it follows that $Z_n^{**} \underset{d^*}{\longrightarrow} Z$. If we can show that Z_n^{**} is uniformly square integrable, Theorem 10.9 shows that $\text{var}\left[Z_n^{**}\right] \to \text{var}[Z] = \boldsymbol{V}_\theta$ as $n \to \infty$. This is shown in the following result, whose proof is presented in Section 10.31.

Theorem 10.12 Under the assumptions of Theorem 10.7, $\widehat{V}_\theta^{\text{boot},\tau} \underset{p^*}{\longrightarrow} \boldsymbol{V}_\theta$.

Theorems 10.11 and 10.12 show that the trimmed bootstrap estimator of variance is consistent for the asymptotic variance in the smooth function model, which includes most econometric estimators. This justifies bootstrap standard errors as consistent estimators for the asymptotic distribution.

An important caveat is that these results critically rely on the trimmed variance estimator. This is a critical caveat as conventional statistical packages (e.g. Stata) calculate bootstrap standard errors using the untrimmed estimator (10.7). Thus there is no guarantee that the reported standard errors are consistent. The untrimmed variance estimator works in the context of Theorem 10.10 and whenever the bootstrap statistic is uniformly square integrable, but not necessarily in general applications.

In practice, it may be difficult to know how to select the trimming sequence τ_n. The rule $\tau_n = O\left(e^{n/8}\right)$ does not provide practical guidance. Instead, it may be useful to think about trimming in terms of percentages of the bootstrap draws. Thus we can set τ_n so that a given small percentage γ_n is trimmed. For theoretical interpretation we would set $\gamma_n \to 0$ as $n \to \infty$. In practice we might set $\gamma_n = 1\%$.