

Econometrics

©Fumio Hayashi

February 16, 2000

Contents

1 Preface	4
1 Finite-Sample Properties of OLS	0
1.1 The Classical Linear Regression Model	1
1.2 The Algebra of Least Squares	11
1.3 Finite-Sample Properties of OLS	21
1.4 Hypothesis Testing under Normality	27
1.5 Relation to Maximum Likelihood	40
1.6 GLS (Generalized Least Squares)	47
1.7 Application: Returns to Scale in Electricity Supply	52
2 Large-Sample Theory	89
2.1 Review of Limit Theorems for Sequences of Random Variables	90
2.2 Fundamental Concepts in Time Series Analysis	98
2.3 Large-Sample Distribution of the OLS Estimator	109
2.4 Hypothesis Testing	117
2.5 Estimating $E(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i')$ Consistently	123
2.6 Implications of Conditional Homoskedasticity	126
2.7 Testing Conditional Homoskedasticity	130
2.8 Estimation with Parameterized Conditional Heteroskedasticity (optional)	133
2.9 Least Squares Projection	137
2.10 Testing for Serial Correlation	141
2.11 Application: Rational Expectations Econometrics	149
2.12 Time Regressions	157
A Asymptotics with Fixed Regressors	162
B Proof of Proposition 2.10	163

3	Single-Equation GMM	190
3.1	Endogeneity Bias: Working's Example	191
3.2	More Examples	196
3.3	The General Formulation	201
3.4	Generalized Method of Moments (GMM) Defined	207
3.5	Large Sample Properties of GMM	211
3.6	Testing Over-identifying Restrictions	219
3.7	Hypothesis Testing by the Likelihood-Ratio Principle	223
3.8	Implications of Conditional Homoskedasticity	227
3.9	Application: Returns from Schooling	237
4	Multiple-Equation GMM	262
4.1	The Multiple-Equation Model	263
4.2	Multiple-Equation GMM Defined	269
4.3	Large Sample Theory	271
4.4	Single-Equation vs. Multiple-Equation Estimation	274
4.5	Special Cases of Multiple-Equation GMM: FIVE, 3SLS, and SUR	278
4.6	Common Coefficients	289
4.7	Application: Interrelated Factor Demands	298
5	Panel Data	326
5.1	The Error-Components Model	327
5.2	The Fixed-Effects Estimator	332
5.3	Unbalanced Panels (optional)	340
5.4	Application: International Differences in Growth Rates	344
A	Distribution of Hausman Statistic	349
6	Serial Correlation	369
6.1	Modelling Serial Correlation: Linear Processes	370
6.2	ARMA Processes	379
6.3	Vector Processes	390
6.4	Estimating Autoregressions	395
6.5	Asymptotics for Sample Means of Serially Correlated Processes	403
6.6	Incorporating Serial Correlation in GMM	409
6.7	Estimation under Conditional Homoskedasticity (optional)	415
6.8	Application: Forward Exchange Rates as Optimal Predictors	420

7	Extremum Estimators	452
7.1	Extremum Estimators	454
7.2	Consistency	463
7.3	Asymptotic Normality	476
7.4	Hypothesis Testing	492
7.5	Numerical Optimization	501
8	Examples of Maximum Likelihood	511
8.1	Qualitative Response (QR) Models	512
8.2	Truncated Regression Models	515
8.3	Censored Regression (Tobit) Models	521
8.4	Multivariate Regressions	525
8.5	FIML	529
8.6	LIML	540
8.7	Serially Correlated Observations	545
9	Unit-Root Econometrics	560
9.1	Modeling Trends	561
9.2	Tools for Unit-Root Econometrics	566
9.3	Dickey-Fuller Tests	574
9.4	Augmented Dickey-Fuller Tests	586
9.5	Which Unit-Root Test to Use?	601
9.6	Application: Purchasing Power Parity	603
10	Cointegration	628
10.1	Cointegrated Systems	630
10.2	Alternative Representations of Cointegrated Systems	638
10.3	Testing the Null of No Cointegration	647
10.4	Inference on Cointegrating Vectors	654
10.5	Application: the Demand for Money in the U.S.	663
A	Partitioned Matrices and Kronecker Products	1

Chapter 1

Preface

This book is designed to serve as the text for a first-year graduate course in econometrics. It has two distinguishing features. First, it covers a full range of techniques with the estimation method called GMM (Generalized Method of Moments) as the organizing principle. I believe that this unified approach is the most efficient way to cover the first-year materials in an accessible yet rigorous manner. Second, most chapters include a section examining in detail original applied articles from such diverse fields in economics as industrial organization, labor, finance, international, and macroeconomics. So the reader will know how to *use* the techniques covered in the chapter and under what conditions they are applicable.

Over the last several years, the lecture notes on which this book is based have been used (successfully, according to student reactions) at University of Pennsylvania, Columbia, Princeton, University of Tokyo, Boston College, and Harvard.

Prerequisites

The reader of this book is assumed to have a working knowledge of the very basics of calculus, probability theory, and linear algebra. An understanding of the following concepts is taken for granted: functions of several variables, partial derivatives, integrals, random variables, joint distributions, independence, unconditional and conditional expectations, variances and covariances of vector random variables, normal distributions, chi-square distributions, matrix multiplication, inverses of matrices, the rank of a matrix, determinants, and positive definite matrices. Any relevant concepts above this level will be introduced as the discussion progresses. Results on partitioned matrices and Kronecker products are collected in an appendix at the end of the book. Prior exposure to undergraduate econometrics is not required, although it would make the reader more appreciative of the book's expositional

efficiency.

Organization of the Book

To understand how the book is organized, it is useful to remember the distinction between a *model* and an *estimation procedure*. The basic premise of econometrics is that economic data (such as postwar U.S. GDP) are a realization of a set of random variables. A model is a family of probability distributions that could possibly have generated the economic data. An estimation procedure is a data-based protocol for choosing from the model a particular distribution that is likely to have generated the data. Most estimation procedures in econometrics are a specialization of the GMM estimation principle. For example, when GMM is applied to a model called the classical linear regression model, the resulting estimation procedure is OLS (Ordinary Least Squares), the most basic estimation procedure in econometrics. This viewpoint is the organizing principle in the first six chapters of the book, where most of the standard estimation procedures are presented.

The book, therefore, could have presented GMM in the very first chapter, but that would rule out the option for the reader to follow a series of topics specific to OLS without getting distracted by GMM. For this reason I chose to use the first two chapters to present the finite-sample and large sample theory of OLS. GMM is presented in Chapter 3 as a generalization of OLS.

A major expositional innovation of the book is to treat multiple-equation estimation procedures — such as SUR (Seemingly Unrelated Regressions), 3SLS (Three-stage Least Squares), the Random-Effects method, covered in Chapter 4, and the Fixed-Effects method covered in Chapter 5 — as special cases of the single-equation GMM of Chapter 3. Therefore, you can derive the statistical properties of those advanced techniques just by suitably specializing the results about single-equation GMM developed in Chapter 3. Chapter 6 completes the book's discussion of GMM by indicating how serial dependence in the error term can be incorporated in GMM.

For some models in econometrics, ML (Maximum Likelihood) is the more natural estimation principle than GMM. ML is covered in Chapters 7 and 8. To make clear the relationship between GMM and ML, the book's discussion of ML starts out in Chapter 7 with an estimation principle called extremum estimators, which includes both ML and GMM as special cases. Applications of ML to various models are covered in Chapter 8.

The book also includes an extensive treatment of time-series analysis. Basic time-series topics are covered in Section 2.2 and in the first half of Chapter 6. That is enough of a prerequisite for the important recent advances in nonstationary time-series analysis, which are covered in the last two chapters of the book.

Designing a Course Out of the Book

Several different courses can be designed based on the book.

- Assuming that the course meets twice for an hour and a half, eight weeks should be enough to cover core theory, which is Chapters 1-4 and 6 (excluding those sections on specific economic applications), Chapter 7 (with proofs and examples skipped), and Chapter 8.
- A 12-week semester course can cover, in addition to the core theory, Chapter 5 and the economic applications included in Chapters 1-6.
- A short (say, six-week) course specializing in GMM estimation in cross-section and panel data would cover Chapters 1-5 (excluding Sections 2.10-2.12 but including Section 2.2). Chapters 7 and 8 (excluding Section 8.7) would add the ML component to the course.
- A short time-series course covering recent developments with economic applications would consist of Chapter 1 (excluding Section 1.7), Chapter 2, Sections 6.1-6.5, Section 6.8, Section 7.1, and Chapters 9 and 10. The applications sections in Chapters 2, 6, 9, and 10 can be dropped if the course focuses on just theory.

Review Questions and Analytical Exercises

The book includes a large number of short questions for review placed at the end of each chapter with lots of hints (and even answers). They can be used to check whether the reader actually understood the materials of the section. On the second reading, if not on the first, you should try to answer them before proceeding. There are several analytical exercises at the end of each chapter. They ask you to prove those results left unproved in the text or else some supplementary results that are useful for their own sake. Unless otherwise indicated, analytical exercises can be skipped without losing continuity.

Empirical Exercises

Each chapter of the book usually has one big empirical exercise. It asks you to replicate the empirical results of the original article discussed in the applications section of the chapter and also estimate various extensions of the article's empirical model using the estimation procedures presented in the chapter. The dataset for estimation, which usually is the same as the one used by the original article, can be downloaded from my website www.e.u-tokyo.ac.jp/hayashi.

To implement the estimation procedure on the dataset, you need to run a statistical package on a computer. There are many statistical packages that are widely used in econometrics. They include: GAUSS (home page: www.aptech.com), MATLAB (www.mathworks.com), TSP (www.tsp.com), RATS (www.estima.com), Stata (www.stata.com), LIMDEP (www.limdep.com), and SAS (www.sas.com). GAUSS and MATLAB are different from the rest in that they are matrix-based languages rather than a collection of procedures. Consider, for example, carrying out OLS with GAUSS or MATLAB. After loading the dataset into the computer's workspace, it takes several lines reflecting the matrix operations of OLS to calculate the OLS estimate and associated statistics (such as R^2). With the other packages, which are procedure-driven and sometimes called "canned packages", those several lines can be replaced by just one-line command invoking the OLS procedure. For example, TSP's OLS command is `OLSQ`.

There are advantages and disadvantages with canned packages. Obviously, it takes far fewer lines to accomplish the same thing, so you can spend less time on programming. On the other hand, procedures in a canned package, which accept data and spit out point estimates and associated statistics, are essentially a black box. Sometimes it is not clear from the documentation of the package how certain statistics are calculated. Although those canned packages mentioned above regularly incorporate new developments in econometrics, the estimation procedure you wish to carry out may not be currently supported by the package, in which case you will be forced to write your own procedures in GAUSS or MATLAB. But it may be a blessing in disguise; actually writing down the underlying matrix operations is a learning experience.

With only a few exceptions, all the calculations needed for the empirical exercises of the book can be carried out with any of the canned packages mentioned above. My recommendation, therefore, is: if you are an economics Ph.D. student planning to write an applied thesis using state-of-the-art estimation procedures or a theoretical thesis proposing new ones, use GAUSS or MATLAB. Otherwise use any of the canned packages mentioned above.

Mathematical Notation

There is no single mathematical notation used by everyone in econometrics. The book's notation follows the most standard, if not universal, practice. Vectors are treated as column vectors and written in bold lowercase letters. Matrices are in bold uppercase letters. The transpose of the matrix \mathbf{A} is denoted by \mathbf{A}' . Scalar variables are (mostly lower case) letters in italics.

Acknowledgements

I acknowledge with gratitude help from the following individuals and institutions. Mark Watson, Dale Jorgenson, Bo Honore, and Serena Ng were kind enough to use early versions of the book as the textbook for their econometrics courses. Comments made by them and their students have been incorporated in this final version. Yuzo Honda read the manuscript and offered helpful suggestions. Naoto Kunitomo, Whitney Newey, Serena Ng, Pierre Perron, Jim Stock, Katsuto Tanaka, Mark Watson, Hal White, and Yoshihiro Yajima took time out to answer my questions and enquiries. Two graduate students at University of Tokyo, Mari Sakudo and Naoki Shimoi, read the entire manuscript to weed out typos. Their effort was underwritten by a grant-in-aid from the Japan Bankers Association. Peter Dougherty, the economics editor at Princeton University Press provided enthusiasm and just the right amount of pressure on me. Jessica Helfand agreed, probably out of friendship, to do the cover design.

For more than five years all my free time went into writing this book. Now, having completed the book, I feel like someone who has just been released from prison. My research suffered, but hopefully the profession hasn't noticed.

Fumio Hayashi

Spring 2000

Chapter 1

Finite-Sample Properties of OLS

Contents

1.1	The Classical Linear Regression Model	1
1.2	The Algebra of Least Squares	11
1.3	Finite-Sample Properties of OLS	21
1.4	Hypothesis Testing under Normality	27
1.5	Relation to Maximum Likelihood	40
1.6	GLS (Generalized Least Squares)	47
1.7	Application: Returns to Scale in Electricity Supply .	52

Abstract

The **Ordinary Least Squares** (OLS) estimator is the most basic estimation procedure in econometrics. This chapter covers the **finite-** or **small-sample properties** of the OLS estimator, that is, the statistical properties of the OLS estimator that are valid for any given sample size. The materials covered in this chapter are entirely standard. The exposition here differs from most other textbooks in its emphasis on the role played by the assumption that the regressors are “strictly exogenous”.

In the final section, we apply the finite-sample theory to the estimation of the cost function using cross-section data on individual firms. The question posed in Nerlove’s (1963) study is of great practical importance: are there increasing returns to scale in electricity supply? If yes, microeconomics tells us that the industry should be regulated. Besides providing you with a hands-on experience of using the techniques to test interesting hypotheses, Nerlove’s paper has a careful discussion of why the OLS is an appropriate estimation procedure in this particular application.

1.1 The Classical Linear Regression Model

In this section we present the assumptions that comprise the classical linear regression model. In the model, the variable in question (called the **dependent variable**, the **regressand**, or more generically the **left-hand (-side) variable**) is related to several other variables (called the **regressors**, the **explanatory variables**, or the **right-hand (-side) variables**). Suppose we observe n values for those variables. Let y_i be the i -th observation of the dependent variable in question and let $(x_{i1}, x_{i2}, \dots, x_{iK})$ be the i -th observation of the K regressors. The **sample** or **data** is a collection of those n observations.

The data in economics cannot be generated by experiments (except in experimental economics), so both the dependent and independent variables have to be treated as random variables, variables whose values are subject to chances. A **model** is a set of restrictions on the joint distribution of the dependent and independent variables. That is, a model is a set of joint distributions satisfying a set of assumptions. The classical regression model is a set of joint distributions satisfying Assumptions 1.1–1.4 stated below.

The Linearity Assumption

The first assumption is that the relationship between the dependent variable and the regressors is linear.

Assumption 1.1 (linearity):

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_i \quad (i = 1, 2, \dots, n), \quad (1.1.1)$$

where β 's are unknown parameters to be estimated, and ε_i is the unobserved error term with certain properties to be specified below.

The part of the right-hand-side involving the regressors, $\beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_K x_{iK}$, is called the **regression** or the **regression function**, and the coefficients (β 's) are called the **regression coefficients**. They represent the marginal and separate effects of the regressors. For example, β_2 represents the change in the dependent variable when the second regressor increases by one unit while other regressors are held constant. In the language of calculus, this can be expressed as: $\partial y_i / \partial x_{i2} = \beta_2$. The linearity implies that the marginal effect does not depend on the level of regressors. The error term represents the part of the dependent variable left unexplained by the regressors.

Example 1.1 (consumption function): The simple consumption function familiar from introductory economics is

$$CON_i = \beta_1 + \beta_2 YD_i + \varepsilon_i, \quad (1.1.2)$$

where CON is consumption and YD is disposable income. If the data are annual aggregate time-series, CON_i and YD_i are aggregate consumption and disposable income for year i . If the data come from a survey of individual households, CON_i is consumption by the i -th household in the cross-section sample of n households. The consumption function can be written as (1.1.1) by setting $y_i = CON_i$, $x_{i1} = 1$ (a constant), and $x_{i2} = YD_i$. The error term ε_i represents other variables besides disposable income that influence consumption. They include those variables — such as financial assets — that might be observable but the researcher decided not to include as regressors, as well as those variables — such as the “mood” of the consumer — that are hard to measure. When the equation has only one non-constant regressor, as here, it is called the **simple regression model**.

The linearity assumption is not as restrictive as it might first seem to you, because the dependent variable and the regressors can be transformations of the variables in question. Consider

Example 1.2 (wage equation): A simplified version of the wage equation routinely estimated in labor economics is

$$\log(WAGE_i) = \beta_1 + \beta_2 S_i + \beta_3 TENURE_i + \beta_4 EXPR_i + \varepsilon_i, \quad (1.1.3)$$

where $WAGE$ = the wage rate for the individual, S = education in years, $TENURE$ = years on the current job, and $EXPR$ = experience in the labor force (i.e., total number of years to date on all the jobs held currently or previously by the individual). The wage equation fits the generic format (1.1.1) with $y_i = \log(WAGE_i)$. The equation is said to be in the **semi-log** form because only the dependent variable is in logs. The equation is derived from the following nonlinear relationship between the level of the wage rate and the regressors:

$$WAGE_i = \exp(\beta_1) \exp(\beta_2 S_i) \exp(\beta_3 TENURE_i) \exp(\beta_4 EXPR_i) \exp(\varepsilon_i). \quad (1.1.4)$$

By taking logs of both sides of (1.1.4) and noting that $\log[\exp(x)] = x$, one obtains (1.1.3). The coefficients in the semi-log form have the interpretation of *percentage changes*, not changes in levels. For example, a value of 0.05 for β_2 implies that an additional year of education has the effect of raising the wage rate by 5%. The difference in the interpretation comes about because the dependent variable is the log wage rate, not the wage rate itself, and the change in logs equals the percentage change in levels.

Certain other forms of nonlinearities can also be accommodated. Suppose, for example, the marginal effect of education tapers off as the level of education gets higher. This can be captured by including in the wage equation the squared term S^2 as an additional regressor in the wage equation. If the coefficient of the squared term is β_5 , the marginal effect of education is

$$\beta_2 + 2\beta_5 S \quad (= \partial \log(WAGE) / \partial S).$$

If β_5 is negative, the marginal effect of education declines with the level of education.

There are, of course, cases of genuine nonlinearity. For example, the relationship (1.1.4) couldn't have been made linear if the error term entered additively rather than multiplicatively:

$$WAGE_i = \exp(\beta_1) \exp(\beta_2 S_i) \exp(\beta_3 TENURE_i) \exp(\beta_4 EXPR_i) + \varepsilon_i.$$

Estimation of nonlinear regression equations such as this will be discussed in Chapter 7.

Matrix Notation

Before stating other assumptions of the classical model, we introduce the vector and matrix notation. The notation will prove useful for stating other assumptions precisely and also for deriving the OLS estimator of $\boldsymbol{\beta}$. Define K -dimensional (column) vectors \mathbf{x}_i and $\boldsymbol{\beta}$ as

$$\underset{(K \times 1)}{\mathbf{x}_i} = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{bmatrix}, \quad \underset{(K \times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}. \quad (1.1.5)$$

By the definition of vector inner products, $\mathbf{x}'_i \boldsymbol{\beta} = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK}$. So the equations in Assumption 1.1 can be written as

$$y_i = \mathbf{x}'_i \boldsymbol{\beta} + \varepsilon_i \quad (i = 1, 2, \dots, n). \quad (1.1.1')$$

Also define

$$\underset{(n \times 1)}{\mathbf{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{(n \times 1)}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \underset{(n \times K)}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1K} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{nK} \end{bmatrix}. \quad (1.1.6)$$

In the vectors and matrices in (1.1.6), there are as many rows as there are observations, with the rows corresponding to the observations. For this reason \mathbf{y} and \mathbf{X}

are sometimes called the **data vector** and the **data matrix**. Since the number of columns of \mathbf{X} equals the number of rows of $\boldsymbol{\beta}$, \mathbf{X} and $\boldsymbol{\beta}$ are conformable and $\mathbf{X}\boldsymbol{\beta}$ is an $n \times 1$ vector. Its i -th element is $\mathbf{x}'_i\boldsymbol{\beta}$. Therefore, Assumption 1.1 can be written compactly as

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{\underbrace{(n \times K)(K \times 1)}_{(n \times 1)}}{\mathbf{X}\boldsymbol{\beta}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}.$$

The Strict Exogeneity Assumption

The next assumption of the classical regression model is

Assumption 1.2 (strict exogeneity):

$$E(\varepsilon_i | \mathbf{X}) = 0 \quad (i = 1, 2, \dots, n). \quad (1.1.7)$$

Here, the expectation (mean) is conditional on the regressors for *all* observations. This point may be made more apparent by writing the assumption without using the data matrix as

$$E(\varepsilon_i | \mathbf{x}_1, \dots, \mathbf{x}_n) = 0 \quad (i = 1, 2, \dots, n).$$

To state the assumption differently, take, for any given observation i , the joint distribution of the $nK + 1$ random variables, $f(\varepsilon_i, \mathbf{x}_1, \dots, \mathbf{x}_n)$, and consider the conditional distribution, $f(\varepsilon_i | \mathbf{x}_1, \dots, \mathbf{x}_n)$. The conditional mean $E(\varepsilon_i | \mathbf{x}_1, \dots, \mathbf{x}_n)$ is in general a nonlinear function of $(\mathbf{x}_1, \dots, \mathbf{x}_n)$. The strict exogeneity assumption says that this function is a constant of zero.¹

Assuming this constant to be zero is not restrictive if the regressors include a constant, because the equation can be rewritten so that the conditional mean of the error term is zero. To see this, suppose that $E(\varepsilon_i | \mathbf{X})$ is μ and $x_{i1} = 1$. The equation can be written as

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_i \\ &= (\beta_1 + \mu) + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + (\varepsilon_i - \mu). \end{aligned}$$

If we redefine β_1 to be $\beta_1 + \mu$ and ε_i to be $\varepsilon_i - \mu$, the conditional mean of the new error term is zero. In virtually all applications, the regressors include a constant term.

¹You should be aware that some authors define the term “strict exogeneity” somewhat differently. For example, in Koopmans and Hood (1953) and Engle, Hendry, and Richards (1983), the regressors are strictly exogenous if \mathbf{x}_i is independent of ε_j for all i, j . This definition is stronger than, but not inconsistent with, our definition of strict exogeneity.

Example 1.3 (continuation of Example 1.1): For the simple regression model of Example 1.1, the strict exogeneity assumption can be written as

$$E(\varepsilon_i \mid YD_1, YD_2, \dots, YD_n) = 0.$$

Since $\mathbf{x}_i = (1, YD_i)'$, you might wish to write the strict exogeneity assumption as

$$E(\varepsilon_i \mid 1, YD_1, 1, YD_2, \dots, 1, YD_n) = 0.$$

But since a constant provides no information, the expectation conditional on

$$(1, YD_1, 1, YD_2, \dots, 1, YD_n)$$

is the same as the expectation conditional on

$$(YD_1, YD_2, \dots, YD_n).$$

Implications of Strict Exogeneity

The strict exogeneity assumption has several implications.

- The *unconditional* mean of the error term is zero, i.e.,

$$E(\varepsilon_i) = 0 \quad (i = 1, 2, \dots, n). \quad (1.1.8)$$

This is because, by the Law of Total Expectations from basic probability theory,² $E[E(\varepsilon_i \mid \mathbf{X})] = E(\varepsilon_i)$.

- If the cross moment $E(xy)$ of two random variables x and y is zero, then we say that x is **orthogonal** to y (or y is orthogonal to x). Under strict exogeneity, the regressors are orthogonal to the error term for *all* observations, i.e.,

$$E(x_{jk}\varepsilon_i) = 0 \quad (i, j = 1, \dots, n; k = 1, \dots, K)$$

or

$$E(\mathbf{x}_{j\cdot}\varepsilon_i) = \begin{bmatrix} E(x_{j1}\varepsilon_i) \\ E(x_{j2}\varepsilon_i) \\ \vdots \\ E(x_{jK}\varepsilon_i) \end{bmatrix} = \begin{matrix} \mathbf{0} \\ (K \times 1) \end{matrix} \quad (\text{for all } i, j). \quad (1.1.9)$$

The proof is a good illustration of the use of properties of conditional expectations and goes as follows.

²The Law of Total Expectation states that $E[E(y \mid \mathbf{x})] = E(y)$.

PROOF. Since x_{jk} is an element of \mathbf{X} , strict exogeneity implies

$$E(\varepsilon_i | x_{jk}) = E[E(\varepsilon_i | \mathbf{X}) | x_{jk}] = 0 \quad (1.1.10)$$

by the Law of Iterated Expectations from probability theory.³ It follows from this that

$$\begin{aligned} E(x_{jk}\varepsilon_i) &= E[E(x_{jk}\varepsilon_i | x_{jk})] \quad (\text{by the Law of Total Expectations}) \\ &= E[x_{jk} E(\varepsilon_i | x_{jk})] \quad (\text{by the linearity of conditional expectations}^4) \\ &= 0. \quad \blacksquare \end{aligned}$$

The point here is that strict exogeneity requires the regressors be orthogonal not only to the error term from the same observation (i.e., $E(x_{ik}\varepsilon_i) = 0$ for all k) but also to the error term from the other observations (i.e., $E(x_{jk}\varepsilon_i) = 0$ for all k and for $j \neq i$).

- Because the mean of the error term is zero, the orthogonality conditions (1.1.9) are equivalent to zero-correlation conditions. This is because

$$\begin{aligned} \text{Cov}(\varepsilon_i, x_{jk}) &= E(x_{jk}\varepsilon_i) - E(x_{jk}) E(\varepsilon_i) \quad (\text{by definition of covariance}) \\ &= E(x_{jk}\varepsilon_i) \quad (\text{since } E(\varepsilon_i) = 0, \text{ see (1.1.8)}) \\ &= 0 \quad (\text{by the orthogonality conditions (1.1.9)}). \end{aligned}$$

In particular, for $i = j$, $\text{Cov}(x_{ik}, \varepsilon_i) = 0$. Therefore, strict exogeneity implies the requirement (familiar to those of you who have studied econometrics before) that the regressors be contemporaneously uncorrelated with the error term.

Strict Exogeneity in Time Series Models

For time-series models where i is time, the implication (1.1.9) of strict exogeneity can be rephrased as: the regressors are orthogonal to the past, current, and future error terms (or equivalently, the error term is orthogonal to the past, current and future regressors). But for most time-series models, this condition (and *a fortiori* strict exogeneity) is not satisfied, so the finite-sample theory based on strict exogeneity to be developed in this section is rarely applicable in time-series contexts. However, as will be shown in the next chapter, the estimator possesses good large-sample properties without strict exogeneity.

³The Law of Iterated Expectation states that $E[E(y | \mathbf{x}, \mathbf{z}) | \mathbf{x}] = E(y | \mathbf{x})$.

⁴The linearity of conditional Expectations states that $E[f(\mathbf{x})y | \mathbf{x}] = f(\mathbf{x})E(y | \mathbf{x})$.

The clearest example of a failure of strict exogeneity is models where the regressors include the **lagged dependent variable**. Consider the simplest such model:

$$y_i = \beta y_{i-1} + \varepsilon_i \quad (i = 1, 2, \dots, n). \quad (1.1.11)$$

This is called the **first-order autoregressive model** (AR(1) for short) (we will study this model more fully in Chapter 6). Suppose, consistent with the spirit of the strict exogeneity assumption, that the regressor for observation i , y_{i-1} , is orthogonal to the error term for i so $E(y_{i-1}\varepsilon_i) = 0$. Then

$$\begin{aligned} E(y_i\varepsilon_i) &= E[(\beta y_{i-1} + \varepsilon_i)\varepsilon_i] \quad (\text{by (1.1.11)}) \\ &= \beta E(y_{i-1}\varepsilon_i) + E(\varepsilon_i^2) \\ &= E(\varepsilon_i^2) \quad (\text{since } E(y_{i-1}\varepsilon_i) = 0 \text{ by hypothesis}). \end{aligned}$$

Therefore, unless the error term is always zero, $E(y_i\varepsilon_i)$ is not zero. But y_i is the regressor for observation $i+1$. Thus, the regressor is not orthogonal to the past error term, which is a violation of strict exogeneity.

Other Assumptions of the Model

The remaining assumptions comprising the classical regression model are the following.

Assumption 1.3 (no multi-collinearity): *The rank of the $n \times K$ data matrix, \mathbf{X} , is K with probability 1.*

Assumption 1.4 (spherical error variance):

$$(\text{homoskedasticity}) \quad E(\varepsilon_i^2 \mid \mathbf{X}) = \sigma^2 > 0 \quad (i = 1, 2, \dots, n),^5 \quad (1.1.12)$$

$$(\text{no correlation between observations}) \quad E(\varepsilon_i\varepsilon_j \mid \mathbf{X}) = 0 \quad (i, j = 1, 2, \dots, n; i \neq j). \quad (1.1.13)$$

To understand Assumption 1.3, recall from matrix algebra that the rank of a matrix equals the number of linearly independent columns of the matrix. The assumption says that none of the K columns of the data matrix \mathbf{X} can be expressed as a linear combination of the other columns of \mathbf{X} . That is, \mathbf{X} is of **full column rank**. Since the K columns cannot be linearly independent if their dimension is less

⁵When a symbol (which here is σ^2) is given to a moment (which here is the second moment $E(\varepsilon_i^2 \mid \mathbf{X})$), by implication the moment is assumed to exist and is finite. We will follow this convention for the rest of this book.

than K , the assumption implies that $n \geq K$, i.e., there must be at least as many observations as there are regressors. The regressors are said to be **(perfectly) multi-collinear** if the assumption is not satisfied. It is easy to see in specific applications when the regressors are multi-collinear and what problems arise.

Example 1.4 (continuation of Example 1.2): If no individuals in the sample ever changed jobs, then $TENURE_i = EXPR_i$ for all i , in violation of the no multi-collinearity assumption. There is evidently no way to distinguish the tenure effect on the wage rate from the experience effect. If we substitute this equality into the wage equation to eliminate $TENURE_i$, the wage equation becomes

$$\log(WAGE_i) = \beta_1 + \beta_2 S_i + (\beta_3 + \beta_4) EXPR_i + \varepsilon_i,$$

which shows that only the sum $\beta_3 + \beta_4$, but not β_3 and β_4 separately, can be estimated.

The homoskedasticity assumption (1.1.12) says that the conditional second moment, which in general is a nonlinear function of \mathbf{X} , is a constant. Thanks to strict exogeneity, this condition can be stated equivalently in more familiar terms. Consider the conditional variance $\text{Var}(\varepsilon_i | \mathbf{X})$. It equals the same constant because

$$\begin{aligned} \text{Var}(\varepsilon_i | \mathbf{X}) &\equiv \text{E}(\varepsilon_i^2 | \mathbf{X}) - \text{E}(\varepsilon_i | \mathbf{X})^2 \quad (\text{by definition of conditional variance}) \\ &= \text{E}(\varepsilon_i^2 | \mathbf{X}) \quad (\text{since } \text{E}(\varepsilon_i | \mathbf{X}) = 0 \text{ by strict exogeneity}). \end{aligned}$$

Similarly, (1.1.13) is equivalent to the requirement that

$$\text{Cov}(\varepsilon_i, \varepsilon_j | \mathbf{X}) = 0 \quad (i, j = 1, 2, \dots, n; i \neq j).$$

That is, in the joint distribution of $(\varepsilon_i, \varepsilon_j)$ conditional on \mathbf{X} , the covariance is zero. In the context of time-series models, (1.1.13) states that there is no **serial correlation** in the error term.

Since the (i, j) element of the $n \times n$ matrix $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$ is $\varepsilon_i\varepsilon_j$, Assumption 1.4 can be written compactly as

$$\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}) = \sigma^2 \mathbf{I}_n. \tag{1.1.14}$$

The discussion of the previous paragraph shows that the assumption can also be written as

$$\text{Var}(\boldsymbol{\varepsilon} | \mathbf{X}) = \sigma^2 \mathbf{I}_n.$$

However, (1.1.14) is the preferred expression, because the more convenient measure of variability is second moments (such as $\text{E}(\varepsilon_i^2 | \mathbf{X})$) rather than variances.

This point will become clearer when we deal with the large sample theory in the next chapter. Assumption 1.4 is sometimes called the **spherical** error variance assumption because the $n \times n$ matrix of second moments (which are also variances and covariances) is proportional to the identity matrix \mathbf{I}_n . This assumption will be relaxed later in this chapter.

The Classical Regression Model for Random Samples

The sample (\mathbf{y}, \mathbf{X}) is a **random sample** if $\{y_i, \mathbf{x}_i\}$ is i.i.d. (independently and identically distributed) across observations. Since by Assumption 1.1 ε_i is a function of (y_i, \mathbf{x}_i) and since (y_i, \mathbf{x}_i) is independent of (y_j, \mathbf{x}_j) for $j \neq i$, ε_i is independent of \mathbf{x}_j for $j \neq i$. So

$$\begin{aligned} E(\varepsilon_i | \mathbf{X}) &= E(\varepsilon_i | \mathbf{x}_i), \\ E(\varepsilon_i^2 | \mathbf{X}) &= E(\varepsilon_i^2 | \mathbf{x}_i), \\ \text{and } E(\varepsilon_i \varepsilon_j | \mathbf{X}) &= E(\varepsilon_i | \mathbf{x}_i) E(\varepsilon_j | \mathbf{x}_j) \quad (\text{for } i \neq j). \end{aligned} \tag{1.1.15}$$

(Proving the last equality in (1.1.15) is a review question.) Therefore, Assumptions 1.2 and 1.4 reduce to

$$\text{Assumption 1.2: } E(\varepsilon_i | \mathbf{x}_i) = 0 \quad (i = 1, 2, \dots, n), \tag{1.1.16}$$

$$\text{Assumption 1.4: } E(\varepsilon_i^2 | \mathbf{x}_i) = \sigma^2 > 0 \quad (i = 1, 2, \dots, n), \tag{1.1.17}$$

The implication of the identical distribution aspect of a random sample is that the joint distribution of $(\varepsilon_i, \mathbf{x}_i)$ does not depend on i . So the *unconditional* second moment $E(\varepsilon_i^2)$ is constant across i (this is referred to as **unconditional homoskedasticity**) and the functional form of the conditional second moment $E(\varepsilon_i^2 | \mathbf{x}_i)$ is the same across i . However, Assumption 1.4 — that the *value* of the conditional second moment is the same across i — doesn't follow. Therefore, Assumption 1.4 remains restrictive for the case of a random sample; without it, the conditional second moment $E(\varepsilon_i^2 | \mathbf{x}_i)$ can differ across i through its possible dependence on \mathbf{x}_i . To emphasize the distinction, the restrictions on the conditional second moments, (1.1.12) and (1.1.17), are referred to as **conditional homoskedasticity**.

“Fixed” Regressors

We have presented the classical linear regression model, treating the regressors as random. This is in contrast to the treatment in most textbooks, where \mathbf{X} is assumed to be “fixed” or deterministic. If \mathbf{X} is fixed, then there is no need to distinguish between the conditional distribution of the error term, $f(\varepsilon_i | \mathbf{x}_1, \dots, \mathbf{x}_n)$, and the unconditional distribution, $f(\varepsilon_i)$, so that Assumptions 1.2 and 1.4 can be written

as

$$\text{Assumption 1.2: } E(\varepsilon_i) = 0 \quad (i = 1, \dots, n), \quad (1.1.18)$$

$$\text{Assumption 1.4: } E(\varepsilon_i^2) = \sigma^2 \quad (i = 1, \dots, n); \quad E(\varepsilon_i \varepsilon_j) = 0 \quad (i, j = 1, \dots, n; i \neq j). \quad (1.1.19)$$

Although it is clearly inappropriate for a non-experimental science like econometrics, the assumption of fixed regressors remains popular because the regression model with fixed \mathbf{X} can be interpreted as a set of statements conditional on \mathbf{X} , allowing us to dispense with “| \mathbf{X} ” from the statements such as Assumptions 1.2 and 1.4 of the model.

However, the economy in the notation comes at a price. It is very easy to miss the point that the error term is being assumed to be uncorrelated with current, past, and future regressors. Also, the distinction between the unconditional and conditional homoskedasticity gets lost if the regressors are deterministic. Throughout this book, the regressors are treated as random and, unless otherwise noted, statements conditional on \mathbf{X} are made explicit by inserting “| \mathbf{X} ”.

Questions for Review

1. (Change in units in the semi-log form) In the wage equation, (1.1.3), of Example 1.2, if *WAGE* is measured in cents rather than in dollars, what difference does it make to the equation? **Hint:**

$$\log(xy) = \log(x) + \log(y).$$

2. Show the last equality in (1.1.15). **Hint:** $E(\varepsilon_i \varepsilon_j | \mathbf{X}) = E[\varepsilon_j E(\varepsilon_i | \mathbf{X}, \varepsilon_j) | \mathbf{X}]$. ε_i is independent of $(\varepsilon_j, \mathbf{x}_1, \dots, \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, \dots, \mathbf{x}_n)$ for $i \neq j$.
3. (Combining linearity and strict exogeneity) Show that Assumptions 1.1 and 1.2 imply

$$E(y_i | \mathbf{X}) = \mathbf{x}'_i \boldsymbol{\beta} \quad (i = 1, 2, \dots, n). \quad (1.1.20)$$

4. (Normally distributed random sample) Consider a random sample on consumption and disposable income, (CON_i, YD_i) ($i = 1, 2, \dots, n$). Suppose the joint distribution of (CON_i, YD_i) (which is the same across i because of the random sample assumption) is normal. Clearly, Assumption 1.3 is satisfied; the rank of \mathbf{X} would be less than K only by pure accident. Show that the

other assumptions, Assumptions 1.1, 1.2, and 1.4, are satisfied. **Hint:** If two random variables, y and x are jointly normally distributed, then the regression function is linear in x , i.e.,

$$E(y | x) = \beta_1 + \beta_2 x,$$

and the conditional variance, $\text{Var}(y | x)$, does not depend on x . Here, the fact that the distribution is the same across i is important; if the distribution differed across i , then β_1 and β_2 could vary across i .

5. (Multi-collinearity for the simple regression model) Show that Assumption 1.3 for the simple regression model is that the non-constant regressor (x_{i2}) is really non-constant (i.e., $x_{i2} \neq x_{j2}$ for some pairs of (i, j) , $i \neq j$, with probability one).
6. (An exercise in conditional and unconditional expectations) Show that Assumptions 1.2 and 1.4 imply:

$$\begin{aligned} \text{Var}(\varepsilon_i) &= \sigma^2 \quad (i = 1, 2, \dots, n) \\ \text{and } \text{Cov}(\varepsilon_i, \varepsilon_j) &= 0 \quad (i \neq j; i, j = 1, 2, \dots, n). \end{aligned} \quad (*)$$

Hint: Strict exogeneity implies $E(\varepsilon_i) = 0$. So (*) is equivalent to:

$$\begin{aligned} E(\varepsilon_i^2) &= \sigma^2 \quad (i = 1, 2, \dots, n) \\ \text{and } E(\varepsilon_i \varepsilon_j) &= 0 \quad (i \neq j; i, j = 1, 2, \dots, n). \end{aligned}$$

1.2 The Algebra of Least Squares

This section describes the computational procedure for obtaining the OLS estimate, \mathbf{b} , of the unknown coefficient vector $\boldsymbol{\beta}$ and introduces a few concepts that derive from \mathbf{b} .

OLS Minimizes the Sum of Squared Residuals

Although we don't observe the error term, we can calculate the value implied by a hypothetical value, $\tilde{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$ as

$$y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}}$$

This is called the **residual** for observation i . From this, form the **sum of squared residuals (SSR)**:

$$SSR(\tilde{\boldsymbol{\beta}}) \equiv \sum_{i=1}^n (y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}})^2 = (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}).$$

This sum is also called the **error sum of squares (ESS)** or the **residual sum of squares (RSS)**. It is a function of $\tilde{\boldsymbol{\beta}}$ because the residual depends on it. The **OLS estimate**, \mathbf{b} , of $\boldsymbol{\beta}$ is the $\tilde{\boldsymbol{\beta}}$ that minimizes this function:

$$\mathbf{b} \equiv \underset{\tilde{\boldsymbol{\beta}}}{\operatorname{argmin}} SSR(\tilde{\boldsymbol{\beta}}). \quad (1.2.1)$$

The relationship between $\boldsymbol{\beta}$ (the unknown coefficient vector), \mathbf{b} (the OLS estimate of it), and $\tilde{\boldsymbol{\beta}}$ (a hypothetical value of $\boldsymbol{\beta}$) is illustrated in Figure 1.1 for $K = 1$. Because $SSR(\tilde{\boldsymbol{\beta}})$ is quadratic in $\tilde{\boldsymbol{\beta}}$, its graph has the U shape. The value of $\tilde{\boldsymbol{\beta}}$ corresponding to the bottom is \mathbf{b} , the OLS estimate. Since it depends on the sample (\mathbf{y}, \mathbf{X}) , the OLS estimate \mathbf{b} is in general different from the true value $\boldsymbol{\beta}$; if \mathbf{b} equals $\boldsymbol{\beta}$, it is by sheer accident.

By having squared residuals in the objective function, this method imposes a heavy penalty on large residuals; the OLS estimate is chosen to prevent large residuals for a few observations at the expense of tolerating relatively small residuals for many other observations. We will see in the next section that this particular criterion brings about some desirable properties for the estimate.

Normal Equations

A sure-fire way of solving the minimization problem is to derive the first-order conditions by setting the partial derivatives equal to zero. To this end we seek a K -dimensional vector of partial derivatives, $\partial SSR(\tilde{\boldsymbol{\beta}})/\partial \tilde{\boldsymbol{\beta}}$.⁶ The task is facilitated by writing $SSR(\tilde{\boldsymbol{\beta}})$ as

$$\begin{aligned} SSR(\tilde{\boldsymbol{\beta}}) &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \quad (\text{since the } i\text{-th element of } \mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} \text{ is } y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}}) \\ &= (\mathbf{y}' - \tilde{\boldsymbol{\beta}}' \mathbf{X}')(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \quad (\text{since } (\mathbf{X}\tilde{\boldsymbol{\beta}})' = \tilde{\boldsymbol{\beta}}' \mathbf{X}') \\ &= \mathbf{y}'\mathbf{y} - \tilde{\boldsymbol{\beta}}' \mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}' \mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}' \mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} \quad (\text{since the scalar } \tilde{\boldsymbol{\beta}}' \mathbf{X}'\mathbf{y} \text{ equals its transpose } \mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &\equiv \mathbf{y}'\mathbf{y} - 2\mathbf{a}'\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}' \mathbf{A}\tilde{\boldsymbol{\beta}} \quad \text{with } \mathbf{a} \equiv \mathbf{X}'\mathbf{y} \text{ and } \mathbf{A} \equiv \mathbf{X}'\mathbf{X}. \end{aligned} \quad (1.2.2)$$

⁶If $g: \mathbb{R}^K \rightarrow \mathbb{R}$ is a scalar-valued function of a K -dimensional vector \mathbf{x} , the derivative of g with respect to \mathbf{x} is a K -dimensional vector whose k -th element is $\partial g(\mathbf{x})/\partial x_k$ where x_k is the k -th element of \mathbf{x} . (This K -dimensional vector is called the **gradient**.) Here, the \mathbf{x} is $\tilde{\boldsymbol{\beta}}$ and the function $g(\mathbf{x})$ is $SSR(\tilde{\boldsymbol{\beta}})$.

The term $\mathbf{y}'\mathbf{y}$ does not depend on $\tilde{\boldsymbol{\beta}}$ and so can be ignored in the differentiation of $SSR(\tilde{\boldsymbol{\beta}})$. Recalling from matrix algebra that

$$\frac{\partial(\mathbf{a}'\tilde{\boldsymbol{\beta}})}{\partial\tilde{\boldsymbol{\beta}}} = \mathbf{a} \quad \text{and} \quad \frac{\partial(\tilde{\boldsymbol{\beta}}'\mathbf{A}\tilde{\boldsymbol{\beta}})}{\partial\tilde{\boldsymbol{\beta}}} = 2\mathbf{A}\tilde{\boldsymbol{\beta}} \quad \text{for } \mathbf{A} \text{ symmetric,}$$

the K -dimensional vector of partial derivatives is

$$\frac{\partial SSR(\tilde{\boldsymbol{\beta}})}{\partial\tilde{\boldsymbol{\beta}}} = -2\mathbf{a} + 2\mathbf{A}\tilde{\boldsymbol{\beta}}.$$

The first-order conditions are obtained by setting this equal to zero. Recalling from (1.2.2) that \mathbf{a} here is $\mathbf{X}'\mathbf{y}$ and \mathbf{A} is $\mathbf{X}'\mathbf{X}$ and rearranging, we can write the first-order conditions as

$$\underset{(K \times K)(K \times 1)}{\mathbf{X}'\mathbf{X}} \mathbf{b} = \mathbf{X}'\mathbf{y}. \quad (1.2.3)$$

Here, we have replaced $\tilde{\boldsymbol{\beta}}$ by \mathbf{b} because the OLS estimate \mathbf{b} is the $\tilde{\boldsymbol{\beta}}$ that satisfies the first-order conditions. These K equations are called the **normal equations**.

The vector of residuals evaluated at $\tilde{\boldsymbol{\beta}} = \mathbf{b}$,

$$\underset{(n \times 1)}{\mathbf{e}} \equiv \mathbf{y} - \mathbf{X}\mathbf{b}, \quad (1.2.4)$$

is called the vector of **OLS residuals**. Its i -th element is $e_i \equiv y_i - \mathbf{x}'_i\mathbf{b}$. Rearranging (1.2.3) gives:

$$\begin{aligned} \mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0} \quad \text{or} \quad \mathbf{X}'\mathbf{e} = \mathbf{0} \quad \text{or} \\ \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \cdot e_i = \mathbf{0} \quad \text{or} \quad \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \cdot (y_i - \mathbf{x}'_i\mathbf{b}) = \mathbf{0}, \end{aligned} \quad (1.2.3')$$

which shows that the normal equations can be interpreted as the sample analogue of the orthogonality conditions $E(\mathbf{x}_i \cdot \varepsilon_i) = \mathbf{0}$. This point will be pursued more fully in subsequent chapters.

To be sure, the first-order conditions are just a necessary condition for minimization and we have to check the second-order condition to make sure that \mathbf{b} achieves the minimum, not the maximum. Those of you who are familiar with the Hessian of a function of several variables⁷ can immediately recognize that the second-order condition is satisfied because (as noted below) $\mathbf{X}'\mathbf{X}$ is positive definite. There is, however, a more direct way to show that \mathbf{b} indeed achieves the minimum. It utilizes the “add-and-subtract” strategy, which is effective when the objective function is quadratic as here. Application of the strategy to the algebra of least squares is left to you as an analytical exercise.

⁷The **Hessian** of $g(\mathbf{x})$ is a square matrix whose (k, ℓ) element is $\partial^2 g(\mathbf{x}) / \partial x_k \partial x_\ell$.

Two Expressions for the OLS Estimator

Thus, we have obtained a system of K linear simultaneous equations in K unknowns in \mathbf{b} . By Assumption 1.3 (no multi-collinearity), the coefficient matrix $\mathbf{X}'\mathbf{X}$ is positive definite (see review question 1 below for a proof) and hence nonsingular. So the normal equations can be solved uniquely for \mathbf{b} by pre-multiplying both sides of (1.2.3) by $(\mathbf{X}'\mathbf{X})^{-1}$:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (1.2.5)$$

Viewed as a function of the sample (\mathbf{y}, \mathbf{X}) , (1.2.5) is sometimes called the OLS **estimator**. For any given sample (\mathbf{y}, \mathbf{X}) , the value of this function is the OLS **estimate**. In this book, as in most other textbooks, the two terms will be used almost interchangeably.

Since $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X}/n)^{-1}\mathbf{X}'\mathbf{y}/n$, the OLS estimator can also be rewritten as

$$\mathbf{b} = \mathbf{S}_{\mathbf{xx}}^{-1} \mathbf{s}_{\mathbf{xy}}, \quad (1.2.5')$$

where

$$\mathbf{S}_{\mathbf{xx}} = \frac{1}{n}\mathbf{X}'\mathbf{X} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \quad (\text{sample average of } \mathbf{x}_i \mathbf{x}_i'), \quad (1.2.6a)$$

$$\mathbf{s}_{\mathbf{xy}} = \frac{1}{n}\mathbf{X}'\mathbf{y} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \cdot y_i \quad (\text{sample average of } \mathbf{x}_i \cdot y_i). \quad (1.2.6b)$$

The data matrix form (1.2.5) is more convenient for developing the finite-sample results, while the sample average form (1.2.5') is the form to be utilized for large-sample theory.

Some More Concepts and Algebra

Having derived the OLS estimator of the coefficient vector, we can define a few related concepts.

- The **fitted value** for observation i is defined as $\hat{y}_i \equiv \mathbf{x}_i' \mathbf{b}$. The vector of fitted value, $\hat{\mathbf{y}}$, equals $\mathbf{X}\mathbf{b}$. Thus the vector of OLS residuals can be written as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$.
- The **projection matrix** \mathbf{P} and the **annihilator** \mathbf{M} . They are defined as

$$\mathbf{P}_{(n \times n)} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad (1.2.7)$$

$$\mathbf{M}_{(n \times n)} \equiv \mathbf{I}_n - \mathbf{P}. \quad (1.2.8)$$

They have the following nifty properties (proving them is a review question):

$$\text{Both } \mathbf{P} \text{ and } \mathbf{M} \text{ are symmetric and idempotent,}^8 \quad (1.2.9)$$

$$\mathbf{P}\mathbf{X} = \mathbf{X} \quad (\text{hence the term } \mathbf{projection} \text{ matrix}), \quad (1.2.10)$$

$$\mathbf{M}\mathbf{X} = \mathbf{0} \quad (\text{hence the term } \mathbf{annihilator}). \quad (1.2.11)$$

Since \mathbf{e} is the residual vector at $\tilde{\boldsymbol{\beta}} = \mathbf{b}$, the sum of squared OLS residuals, SSR , equals $\mathbf{e}'\mathbf{e}$. It can further be written as

$$SSR = \mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}. \quad (1.2.12)$$

(Proving this is a review question.) This expression, relating SSR to the true error term $\boldsymbol{\varepsilon}$, will be useful later on.

- The OLS estimate of σ^2 (the variance of the error term), denoted s^2 , is the sum of squared residuals divided by $n - K$:

$$s^2 \equiv \frac{SSR}{n - K} = \frac{\mathbf{e}'\mathbf{e}}{n - K}. \quad (1.2.13)$$

(The definition presumes that $n > K$; otherwise s^2 is not well-defined.) As will be shown in Proposition 1.2 below, dividing the sum of squared residuals by $n - K$ (called the **degrees of freedom**) rather than by n (the sample size) makes this estimate unbiased for σ^2 . The intuitive reason is that K parameters ($\boldsymbol{\beta}$) have to be estimated before obtaining the residual vector \mathbf{e} used to calculate s^2 . More specifically, \mathbf{e} has to satisfy the K normal equations (1.2.3'), which limits the variability of the residual.

- The square root of s^2 , s , is called the the **standard error of the regression (SER for short) or standard error of the equation (SEE)**. It is an estimate of the standard deviation of the error term.
- The **sampling error** is defined as $\mathbf{b} - \boldsymbol{\beta}$. It too can be related to $\boldsymbol{\varepsilon}$ as follows.

$$\begin{aligned} \mathbf{b} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta} \quad (\text{by (1.2.5)}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} \quad (\text{since } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ by Assumption 1.1}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}. \end{aligned} \quad (1.2.14)$$

⁸A square matrix \mathbf{A} is said to be **idempotent** if $\mathbf{A} = \mathbf{A}^2$.

- **Uncentered R^2 .** One measure of the variability of the dependent variable is the sum of squares, $\sum y_i^2 = \mathbf{y}'\mathbf{y}$. Because the OLS residual is chosen to satisfy the normal equations, we have the following decomposition of $\mathbf{y}'\mathbf{y}$:

$$\begin{aligned}
 \mathbf{y}'\mathbf{y} &= (\hat{\mathbf{y}} + \mathbf{e})'(\hat{\mathbf{y}} + \mathbf{e}) \quad (\text{since } \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}) \\
 &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\hat{\mathbf{y}}'\mathbf{e} + \mathbf{e}'\mathbf{e} \\
 &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + 2\mathbf{b}'\mathbf{X}'\mathbf{e} + \mathbf{e}'\mathbf{e} \quad (\text{since } \hat{\mathbf{y}} \equiv \mathbf{X}\mathbf{b}) \\
 &= \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e} \quad (\text{since } \mathbf{X}'\mathbf{e} = \mathbf{0} \text{ by the normal equations; see (1.2.3')}).
 \end{aligned}
 \tag{1.2.15}$$

The **uncentered R^2** is defined as

$$R_{uc}^2 \equiv 1 - \frac{\mathbf{e}'\mathbf{e}}{\mathbf{y}'\mathbf{y}}. \tag{1.2.16}$$

Because of the decomposition (1.2.15), this equals

$$\frac{\hat{\mathbf{y}}'\hat{\mathbf{y}}}{\mathbf{y}'\mathbf{y}}.$$

Since both $\hat{\mathbf{y}}'\hat{\mathbf{y}}$ and $\mathbf{e}'\mathbf{e}$ are nonnegative, $0 \leq R_{uc}^2 \leq 1$. Thus the uncentered R^2 has the interpretation of the fraction of the variation of the dependent variable that is attributable to the variation in the explanatory variables. The closer the fitted value tracks the dependent variable, the closer is the uncentered R^2 to one.

- **(centered) R^2 , the coefficient of determination.** If the only regressor is a constant (so that $K = 1$ and $x_{i1} = 1$), then it is easy to see from (1.2.5) that \mathbf{b} equals \bar{y} , the sample mean of the dependent variable, which means that $\hat{y}_i = \bar{y}$ for all i , $\hat{\mathbf{y}}'\hat{\mathbf{y}}$ in (1.2.15) equals $n\bar{y}^2$, and $\mathbf{e}'\mathbf{e}$ equals $\sum_i (y_i - \bar{y})^2$. If the regressors also include non-constant variables, then it can be shown (the proof is left as an analytical exercise) that $\sum_i (y_i - \bar{y})^2$ is decomposed as

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n e_i^2 \quad \text{with } \bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i. \tag{1.2.17}$$

The **coefficient of determination**, R^2 , is defined as

$$R^2 \equiv 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \tag{1.2.18}$$

Because of the decomposition (1.2.17), this R^2 equals

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

Therefore, provided that the regressors include a constant so that the decomposition (1.2.17) is valid, $0 \leq R^2 \leq 1$. Thus this R^2 as defined in (1.2.18) is a measure of the explanatory power of the non-constant regressors.

If the regressors don't include a constant but if (as some regression software packages do) you nevertheless calculate R^2 by the formula (1.2.18), then the R^2 can be negative. This is because, without the benefit of an intercept, the regression could do worse than the sample mean in terms of tracking the dependent variable. On the other hand, some other regression packages (notably STATA) switch to the formula (1.2.16) for the R^2 when a constant is not included, in order to avoid negative values for the R^2 . This is a mixed blessing. Suppose that the regressors don't include a constant but that a linear combination of the regressors equals a constant. This occurs if, for example, the intercept is replaced by seasonal dummies.⁹ The regression is essentially the same when one of the regressors in the linear combination is replaced by a constant. Indeed, one should obtain the same vector of fitted values. But if the formula for the R^2 is (1.2.16) for regressions without a constant and (1.2.18) for those with a constant, the calculated R^2 changes (actually declines, see Review Question 7 below) after the replacement by a constant.

Influential Analysis (optional)

Since the method of least squares seeks to prevent a few large residuals at the expense of incurring many relatively small residuals, only a few observations can be extremely influential in the sense that dropping those from the sample changes some elements of \mathbf{b} substantially. There is a systematic way to find those **influential observations**.¹⁰ Let $\mathbf{b}^{(i)}$ be the OLS estimate of β that would be obtained if OLS were used on a sample from which the i -th observation was omitted. The key equation is

$$\mathbf{b}^{(i)} - \mathbf{b} = -\left(\frac{1}{1 - p_i}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \cdot e_i, \quad (1.2.19)$$

where \mathbf{x}_i as before is the i -th row of \mathbf{X} , e_i is the OLS residual for observation i and p_i is defined as

$$p_i \equiv \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i, \quad (1.2.20)$$

which is the i -th diagonal element of the projection matrix \mathbf{P} . (Proving (1.2.19) would be a good exercise in matrix algebra but we won't do it here.) It is easy to

⁹Dummy variables will be introduced in the empirical exercise to this chapter.

¹⁰See Krasker, Kuh, and Welsch (1983) for more details.

Table 1.1: Influential Analysis

Country	GDP/worker growth	Equipment/ GDP	Residual	p_i	(1.2.19) for β_1	(1.2.19) for β_2
Botswana	0.0676	0.1310	0.0119	0.7196	0.0104	-0.3124
Cameroon	0.0458	0.0415	0.0233	0.0773	-0.0021	0.0045
Ethiopia	0.0094	0.0212	-0.0056	0.1193	0.0010	-0.0119
India	0.0115	0.0278	-0.0059	0.0980	0.0009	-0.0087
Indonesia	0.0345	0.0221	0.0192	0.1160	-0.0034	0.0394
Ivory Coast	0.0278	0.0243	0.0117	0.1084	-0.0019	0.0213
Kenya	0.0146	0.0462	-0.0096	0.0775	0.0007	0.0023
Madagascar	-0.0102	0.0219	-0.0254	0.1167	0.0045	-0.0527
Malawi	0.0153	0.0361	-0.0052	0.0817	0.0006	-0.0036
Mali	0.0044	0.0433	-0.0188	0.0769	0.0016	-0.0006
Pakistan	0.0295	0.0263	0.0126	0.1022	-0.0020	0.0205
Tanzania	0.0184	0.0860	-0.0206	0.2281	-0.0021	0.0952
Thailand	0.0341	0.0395	0.0123	0.0784	-0.0012	0.0047

show (see Review Question 7 of Section 1.3) that

$$0 \leq p_i \leq 1 \quad \text{and} \quad \sum_{i=1}^n p_i = K. \quad (1.2.21)$$

So p_i equals K/n on average.

To illustrate the use of (1.2.19) in a specific example, consider the relationship between equipment investment and economic growth for the world's poorest countries between 1960 and 1985. Figure 1.2 plots the average annual GDP per worker growth between 1960 and 85 against the ratio of equipment investment to GDP over the same period for 13 countries whose GDP per worker in 1965 is less than 10% of that of the U.S.¹¹ It is clear visually from the plot that the position of the estimated regression line would depend very much on the single outlier (Botswana). Indeed, if Botswana is dropped from the sample, the estimated slope coefficient drops from 0.37 to 0.058. In the present case of simple regression, it is easy to spot outliers by visually inspecting the plot such as Figure 1.2. This eyeballing strategy would not work if there were more than one non-constant regressors. Analysis based on formula (1.2.19) is not restricted to simple regressions. Table 1.1 displays the data

¹¹The data are from the Penn World Table, reprinted in De Long and Summers (1991). To their credit, their analysis is based on the whole sample of 61 countries.

along with the OLS residuals and the values of p_i , and (1.2.19) for each observation. Botswana's p_i of 0.7196 is well above the average of 0.154 ($= K/n = 2/13$) and is highly **influential**, as the last two columns of the table indicates. Note that we could not have detected the influential observation by looking at the residuals, which is not surprising because the algebra of least squares is designed to avoid large residuals at the expense of many small residuals for other observations.

What to do with influential observations? It depends. If the influential observations satisfy the regression model, they provide valuable information about the regression function unavailable from the rest of the sample and should definitely be kept in the sample. But more probable is that the influential observations are untypical of the rest of the sample because they don't satisfy the model. In this case they should definitely be dropped from the sample. For the example just examined, there was a world-wide growth in the demand for diamonds, Botswana's main export, and production of diamonds requires heavy investment in drilling equipments. If the reason for us to expect an association between growth and equipment investment is the beneficial effect on productivity of the introduction of new technologies through equipment, then Botswana, whose high GDP growth is demand-driven, should be dropped from the sample.

A Note on the Computation of OLS Estimates¹²

So far, we have focused on the conceptual aspects of the algebra of least squares. But for applied researchers who actually calculate OLS estimates using digital computers, it is important to be aware of a certain aspect of digital computing in order to avoid the risk of obtaining unreliable estimates without knowing it. The source of a potential problem is that the computer approximates real numbers by so-called **floating-point numbers**. When an arithmetic operation involves both very large numbers and very small numbers, floating-point calculation can produce inaccurate results. This is relevant in the computation of OLS estimates when the regressors greatly differ in magnitude. For example, one of the regressors may be the interest rate stated as a fraction and another regressor may be U.S. GDP in dollars. The matrix $\mathbf{X}'\mathbf{X}$ will then contain both very small and very large numbers, and the arithmetic operation of inverting this matrix by the digital computer will produce unreliable results.

A simple solution to this problem is to choose the units of measurement so that the regressors are similar in magnitude. For example, state the interest rate in percents and U.S. GDP in trillion dollars. This sort of care would prevent the problem

¹²A fuller treatment of this topic can be found in Section 1.5 of Davidson and MacKinnon (1993).

most of the time. A more systematic transformation of the \mathbf{X} matrix is to subtract the sample means of all regressors and divide by the sample standard deviations before forming $\mathbf{X}'\mathbf{X}$ (and adjust the OLS estimates to undo the transformation). Most OLS programs (such as TSP) take a more sophisticated transformation of the \mathbf{X} matrix (called the **QR decomposition**) to produce accurate results.

Questions for Review

1. Prove that $\mathbf{X}'\mathbf{X}$ is positive definite if \mathbf{X} is of full column rank. **Hint:** What needs to be shown is that $\mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c} > 0$ for $\mathbf{c} \neq \mathbf{0}$. Define $\mathbf{z} \equiv \mathbf{X}\mathbf{c}$. Then $\mathbf{c}'\mathbf{X}'\mathbf{X}\mathbf{c} = \mathbf{z}'\mathbf{z} = \sum_{k=1}^K z_k^2$. If \mathbf{X} is of full column rank, then $\mathbf{z} \neq \mathbf{0}$ for any $\mathbf{c} \neq \mathbf{0}$.
2. Verify that $\mathbf{X}'\mathbf{X}/n = \frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i'$ and $\mathbf{X}'\mathbf{y}/n = \frac{1}{n} \sum_i \mathbf{x}_i \cdot y_i$ as in (1.2.6). **Hint:** The (k, ℓ) element of $\mathbf{X}'\mathbf{X}$ is $\sum_i x_{ik} x_{i\ell}$.
3. (OLS estimator for the simple regression model) In the simple regression model, $K = 2$ and $x_{i1} = 1$. Show that

$$\mathbf{S}_{\mathbf{xx}} = \begin{bmatrix} 1 & \bar{x}_2 \\ \bar{x}_2 & \frac{1}{n} \sum_{i=1}^n x_{i2}^2 \end{bmatrix}, \quad \mathbf{S}_{\mathbf{xy}} = \begin{bmatrix} \bar{y} \\ \frac{1}{n} \sum_{i=1}^n x_{i2} y_i \end{bmatrix}$$

where

$$\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i \quad \text{and} \quad \bar{x}_2 \equiv \frac{1}{n} \sum_{i=1}^n x_{i2}.$$

Show that

$$b_2 = \frac{\frac{1}{n} \sum_{i=1}^n (x_{i2} - \bar{x}_2)(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2} \quad \text{and} \quad b_1 = \bar{y} - \bar{x}_2 b_2.$$

(You may recognize that the denominator of the expression for b_2 as the sample variance of the non-constant regressor and the numerator as the sample covariance between the non-constant regressor and the dependent variable.)

Hint:

$$\frac{1}{n} \sum_{i=1}^n x_{i2}^2 - (\bar{x}_2)^2 = \frac{1}{n} \sum_{i=1}^n (x_{i2} - \bar{x}_2)^2$$

and

$$\frac{1}{n} \sum_{i=1}^n x_{i2} y_i - \bar{x}_2 \bar{y} = \frac{1}{n} \sum_{i=1}^n (x_{i2} - \bar{x}_2)(y_i - \bar{y}).$$

You can take (1.2.5') and use the brute force of matrix inversion. Alternatively, write down the two normal equations. The first normal equation is $b_1 = \bar{y} - \bar{x}_2 b_2$. Substitute this into the second normal equation to eliminate b_1 and then solve for b_2 .

4. Show (1.2.9)–(1.2.11). **Hint:** They should easily follow from the definition of \mathbf{P} and \mathbf{M} .
5. (Matrix algebra of fitted values and residuals) Show the following.
 - (a) $\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$, $\mathbf{e} = \mathbf{M}\mathbf{y} = \mathbf{M}\boldsymbol{\varepsilon}$. **Hint:** Use (1.2.5).
 - (b) (1.2.12), namely $SSR = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$
6. (Change in units and R^2) Does a change in the unit of measurement for the dependent variable change R^2 ? A change in the unit of measurement for the regressors? **Hint:** Check whether the change affects the denominator and the numerator in the definition for R^2 .
7. (Relation between R_{uc}^2 and R^2) Show that

$$1 - R^2 = \left(1 + \frac{n \cdot \bar{y}^2}{\sum_{i=1}^n (y_i - \bar{y})^2}\right) (1 - R_{uc}^2).$$

Hint: Use (1.2.16), (1.2.18), and the identity $\sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n \cdot \bar{y}^2$.

8. Show that

$$R_{uc}^2 = \frac{\mathbf{y}'\mathbf{P}\mathbf{y}}{\mathbf{y}'\mathbf{y}}.$$

9. (Computation of the statistics) Verify that \mathbf{b} , SSR , s^2 , and R^2 can be calculated from the following sample averages: \mathbf{S}_{xx} , \mathbf{s}_{xy} , $\mathbf{y}'\mathbf{y}/n$, and \bar{y} . (If the regressors include a constant, then \bar{y} is the element of \mathbf{s}_{xy} corresponding to the constant.) Therefore, those sample averages need to be computed just once in order to obtain the regression coefficients and related statistics.

1.3 Finite-Sample Properties of OLS

Having derived the OLS estimator, we now examine its finite-sample properties, namely the characteristics of the distribution of the estimator that are valid for any given sample size n .

Finite-Sample Distribution of \mathbf{b}

Proposition 1.1 (finite-sample properties of the OLS estimator of β):

- (a) (unbiasedness) Under Assumptions 1.1–1.3, $E(\mathbf{b} \mid \mathbf{X}) = \beta$.
- (b) (expression for the variance) Under Assumptions 1.1–1.4, $\text{Var}(\mathbf{b} \mid \mathbf{X}) = \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$.
- (c) (Gauss-Markov Theorem) Under Assumptions 1.1–1.4, the OLS estimator is **efficient** in the class of linear unbiased estimators. That is, for any unbiased estimator $\hat{\beta}$ that is linear in \mathbf{y} , $\text{Var}(\hat{\beta} \mid \mathbf{X}) \geq \text{Var}(\mathbf{b} \mid \mathbf{X})$ in the matrix sense.¹³
- (d) Under Assumptions 1.1–1.4, $\text{Cov}(\mathbf{b}, \mathbf{e} \mid \mathbf{X}) = \mathbf{0}$, where $\mathbf{e} \equiv \mathbf{y} - \mathbf{X}\mathbf{b}$.

Before plunging into the proof, let us be clear about what this proposition means.

- The matrix inequality in part (c) says that the $K \times K$ matrix $\text{Var}(\hat{\beta} \mid \mathbf{X}) - \text{Var}(\mathbf{b} \mid \mathbf{X})$ is positive semi-definite, so

$$\mathbf{a}'[\text{Var}(\hat{\beta} \mid \mathbf{X}) - \text{Var}(\mathbf{b} \mid \mathbf{X})]\mathbf{a} \geq 0 \quad \text{or} \quad \mathbf{a}'\text{Var}(\hat{\beta} \mid \mathbf{X})\mathbf{a} \geq \mathbf{a}'\text{Var}(\mathbf{b} \mid \mathbf{X})\mathbf{a}$$

for any K -dimensional vector \mathbf{a} . In particular, consider a special vector whose elements are all 0 except for the k -th element which is 1. For this particular \mathbf{a} , the quadratic form $\mathbf{a}'\mathbf{A}\mathbf{a}$ picks up the (k, k) element of \mathbf{A} . But the (k, k) element of $\text{Var}(\hat{\beta} \mid \mathbf{X})$, for example, is $\text{Var}(\hat{\beta}_k \mid \mathbf{X})$ where $\hat{\beta}_k$ is the k -th element of $\hat{\beta}$. Thus the matrix inequality in (c) implies

$$\text{Var}(\hat{\beta}_k \mid \mathbf{X}) \geq \text{Var}(b_k \mid \mathbf{X}) \quad (k = 1, 2, \dots, K). \quad (1.3.1)$$

That is, for any regression coefficient, the variance of the OLS estimator is no larger than that of any other linear unbiased estimator.

- As clear from (1.2.5), the OLS estimator is linear in \mathbf{y} . There are many other estimators of β that are linear and unbiased (you will be asked to provide one in a review question below). The Gauss-Markov Theorem says that the OLS estimator is efficient in the sense that its conditional variance matrix $\text{Var}(\mathbf{b} \mid \mathbf{X})$ is smallest among linear unbiased estimators. For this reason the OLS estimator is called **BLUE** (Best Linear Unbiased Estimator).

¹³Let \mathbf{A} and \mathbf{B} be two square matrices of the same size. We say that $\mathbf{A} \geq \mathbf{B}$ if $\mathbf{A} - \mathbf{B}$ is positive semi-definite. A $K \times K$ matrix \mathbf{C} is said to be positive semi-definite (or nonnegative definite) if $\mathbf{x}'\mathbf{C}\mathbf{x} \geq 0$ for all K -dimensional vector \mathbf{x} .

- The OLS estimator \mathbf{b} is a function of the sample (\mathbf{y}, \mathbf{X}) . Since (\mathbf{y}, \mathbf{X}) are random, so is \mathbf{b} . Now imagine that we fix \mathbf{X} at some given value, calculate \mathbf{b} for all samples corresponding to all possible realizations of \mathbf{y} , and take the average of \mathbf{b} (the Monte Carlo exercise to this chapter will ask you to actually do this). This average is the (population) conditional mean $E(\mathbf{b} | \mathbf{X})$. Part (a) (unbiasedness) says that this average equals the true value β .
- There is another notion of unbiasedness that is weaker than the unbiasedness of part (a). By the Law of Total Expectations, $E[E(\mathbf{b} | \mathbf{X})] = E(\mathbf{b})$. So (a) implies

$$E(\mathbf{b}) = \beta. \quad (1.3.2)$$

This says: if we calculated \mathbf{b} for all possible different samples, differing not only in \mathbf{y} but also in \mathbf{X} , the average would be the true value. This unconditional statement is probably more relevant in economics because samples do differ in both \mathbf{y} and \mathbf{X} . The import of the conditional statement (a) is that it implies the unconditional statement (1.3.2) which is more relevant.

- The same holds for the conditional statement (c) about the variance. A review question below asks you to show that statements (a) and (b) imply

$$\text{Var}(\hat{\beta}) \geq \text{Var}(\mathbf{b}) \quad (1.3.3)$$

where $\hat{\beta}$ is any linear unbiased estimator (so that $E(\hat{\beta} | \mathbf{X}) = \beta$).

We now go through the proof of this important result. The proof may look lengthy to you. If so, it is only because the proof records every step, however easy. In the first reading, you can skip the proof of part (c). Proof of (d) is a review question.

PROOF.

- (a) (Proof that $E(\mathbf{b} | \mathbf{X}) = \beta$) $E(\mathbf{b} - \beta | \mathbf{X}) = \mathbf{0}$ whenever $E(\mathbf{b} | \mathbf{X}) = \beta$. So we prove the former. By the expression for the sampling error (1.2.14), $\mathbf{b} - \beta = \mathbf{A}\varepsilon$ where \mathbf{A} here is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. So

$$E(\mathbf{b} - \beta | \mathbf{X}) = E(\mathbf{A}\varepsilon | \mathbf{X}) = \mathbf{A} E(\varepsilon | \mathbf{X}).$$

Here, the second equality holds by the linearity of conditional expectations; \mathbf{A} is a function of \mathbf{X} and so can be treated as if non-random. Since $E(\varepsilon | \mathbf{X}) = \mathbf{0}$, the last expression is zero.

(b) (Proof that $\text{Var}(\mathbf{b} \mid \mathbf{X}) = \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$)

$$\begin{aligned}
\text{Var}(\mathbf{b} \mid \mathbf{X}) &= \text{Var}(\mathbf{b} - \boldsymbol{\beta} \mid \mathbf{X}) \quad (\text{since } \boldsymbol{\beta} \text{ is not random}) \\
&= \text{Var}(\mathbf{A}\boldsymbol{\varepsilon} \mid \mathbf{X}) \quad (\text{by (1.2.14) and } \mathbf{A} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \mathbf{A} \text{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})\mathbf{A}' \quad (\text{since } \mathbf{A} \text{ is a function of } \mathbf{X}) \\
&= \mathbf{A} \text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X})\mathbf{A}' \quad (\text{by Assumption 1.2}) \\
&= \mathbf{A}(\sigma^2\mathbf{I}_n)\mathbf{A}' \quad (\text{by Assumption 1.4, see (1.1.14)}) \\
&= \sigma^2\mathbf{A}\mathbf{A}' \\
&= \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1} \quad (\text{since } \mathbf{A}\mathbf{A}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}).
\end{aligned}$$

(c) (Gauss-Markov) Since $\widehat{\boldsymbol{\beta}}$ is linear in \mathbf{y} , it can be written as $\widehat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{y}$ for some matrix \mathbf{C} which possibly is a function of \mathbf{X} . Let $\mathbf{D} \equiv \mathbf{C} - \mathbf{A}$ or $\mathbf{C} = \mathbf{D} + \mathbf{A}$ where $\mathbf{A} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{D} + \mathbf{A})\mathbf{y} \\
&= \mathbf{D}\mathbf{y} + \mathbf{A}\mathbf{y} \\
&= \mathbf{D}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) + \mathbf{b} \quad (\text{since } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{ and } \mathbf{A}\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{b}) \\
&= \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{b}.
\end{aligned}$$

Taking the conditional expectation of both sides, we obtain

$$\text{E}(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) = \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \text{E}(\mathbf{D}\boldsymbol{\varepsilon} \mid \mathbf{X}) + \text{E}(\mathbf{b} \mid \mathbf{X}).$$

Since both \mathbf{b} and $\widehat{\boldsymbol{\beta}}$ are unbiased and since $\text{E}(\mathbf{D}\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{D}\text{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$, it follows that $\mathbf{D}\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$. For this to be true for any given $\boldsymbol{\beta}$, it is necessary that $\mathbf{D}\mathbf{X} = \mathbf{0}$. So $\widehat{\boldsymbol{\beta}} = \mathbf{D}\boldsymbol{\varepsilon} + \mathbf{b}$ and

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \mathbf{D}\boldsymbol{\varepsilon} + (\mathbf{b} - \boldsymbol{\beta}) \\
&= (\mathbf{D} + \mathbf{A})\boldsymbol{\varepsilon} \quad (\text{by (1.2.14)}).
\end{aligned}$$

So

$$\begin{aligned}
\text{Var}(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) &= \text{Var}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \mid \mathbf{X}) \\
&= \text{Var}[(\mathbf{D} + \mathbf{A})\boldsymbol{\varepsilon} \mid \mathbf{X}] \\
&= (\mathbf{D} + \mathbf{A}) \text{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})(\mathbf{D}' + \mathbf{A}') \quad (\text{since both } \mathbf{D} \text{ and } \mathbf{A} \text{ are functions of } \mathbf{X}) \\
&= \sigma^2 \cdot (\mathbf{D} + \mathbf{A})(\mathbf{D}' + \mathbf{A}') \quad (\text{since } \text{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}_n) \\
&= \sigma^2 \cdot (\mathbf{D}\mathbf{D}' + \mathbf{A}\mathbf{D}' + \mathbf{D}\mathbf{A}' + \mathbf{A}\mathbf{A}').
\end{aligned}$$

But $\mathbf{DA}' = \mathbf{DX}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$ since $\mathbf{DX} = \mathbf{0}$. Also, $\mathbf{AA}' = (\mathbf{X}'\mathbf{X})^{-1}$ as shown in (b). So

$$\begin{aligned}\text{Var}(\widehat{\boldsymbol{\beta}} \mid \mathbf{X}) &= \sigma^2 \cdot [\mathbf{DD}' + (\mathbf{X}'\mathbf{X})^{-1}] \\ &\geq \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1} \quad (\text{since } \mathbf{DD}' \text{ is positive semi-definite}) \\ &= \text{Var}(\mathbf{b} \mid \mathbf{X}) \quad (\text{by (b)}). \blacksquare\end{aligned}$$

It should be emphasized that the strict exogeneity assumption (Assumption 1.2) is critical for proving unbiasedness. Anything short of strict exogeneity won't do. For example, it is not enough to assume that $E(\varepsilon_i \mid \mathbf{x}_i) = 0$ for all i or that $E(\mathbf{x}_i \cdot \varepsilon_i) = \mathbf{0}$ for all i . We have noted in Section 1.1 that most time-series models don't satisfy strict exogeneity even if they satisfy weaker conditions such as the orthogonality condition $E(\mathbf{x}_i \cdot \varepsilon_i) = \mathbf{0}$. It follows that for those models the OLS estimator is not unbiased.

Finite-Sample Properties of s^2

We defined the OLS estimator of σ^2 in (1.2.13). It, too, is unbiased.

Proposition 1.2 (Unbiasedness of s^2): *Under Assumptions 1.1–1.4, $E(s^2 \mid \mathbf{X}) = \sigma^2$ (and hence $E(s^2) = \sigma^2$), provided $n > K$ (so that s^2 is well-defined).*

We can prove this proposition easily by the use of the trace operator.¹⁴

PROOF. Since $s^2 = \mathbf{e}'\mathbf{e}/(n - K)$, the proof amounts to showing that $E(\mathbf{e}'\mathbf{e} \mid \mathbf{X}) = (n - K)\sigma^2$. As shown in (1.2.12), $\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ where \mathbf{M} is the annihilator. The proof consists of proving two properties: (1) $E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2 \cdot \text{trace}(\mathbf{M})$, and (2) $\text{trace}(\mathbf{M}) = n - K$.

(1) (Proof that $E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2 \cdot \text{trace}(\mathbf{M})$) Since $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} = \sum_{i=1}^n \sum_{j=1}^n m_{ij} \varepsilon_i \varepsilon_j$ (this is just writing out the quadratic form $\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$), we have

$$\begin{aligned}E(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon} \mid \mathbf{X}) &= \sum_{i=1}^n \sum_{j=1}^n m_{ij} E(\varepsilon_i \varepsilon_j \mid \mathbf{X}) \\ &\quad (\text{Because } m_{ij} \text{'s are functions of } \mathbf{X}, E(m_{ij} \varepsilon_i \varepsilon_j \mid \mathbf{X}) = m_{ij} E(\varepsilon_i \varepsilon_j \mid \mathbf{X})) \\ &= \sum_{i=1}^n m_{ii} \sigma^2 \quad (\text{since } E(\varepsilon_i \varepsilon_j \mid \mathbf{X}) = 0 \text{ for } i \neq j \text{ by Assumption 1.4}) \\ &= \sigma^2 \sum_{i=1}^n m_{ii} \\ &= \sigma^2 \cdot \text{trace}(\mathbf{M}).\end{aligned}$$

¹⁴The **trace** of a square matrix \mathbf{A} is the sum of the diagonal elements of \mathbf{A} : $\text{trace}(\mathbf{A}) = \sum_i a_{ii}$.

(2) (Proof that $\text{trace}(\mathbf{M}) = n - K$)

$$\begin{aligned}\text{trace}(\mathbf{M}) &= \text{trace}(\mathbf{I}_n - \mathbf{P}) \quad (\text{since } \mathbf{M} \equiv \mathbf{I}_n - \mathbf{P}; \text{ see (1.2.8)}) \\ &= \text{trace}(\mathbf{I}_n) - \text{trace}(\mathbf{P}) \quad (\text{fact: the trace operator is linear.}) \\ &= n - \text{trace}(\mathbf{P}),\end{aligned}$$

and

$$\begin{aligned}\text{trace}(\mathbf{P}) &= \text{trace}[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] \quad (\text{since } \mathbf{P} \equiv \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'; \text{ see (1.2.7)}) \\ &= \text{trace}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}] \quad (\text{fact: } \text{trace}(\mathbf{AB}) = \text{trace}(\mathbf{BA})) \\ &= \text{trace}(\mathbf{I}_K) = K.\end{aligned}$$

So $\text{trace}(\mathbf{M}) = n - K$. ■

Estimate of $\text{Var}(\mathbf{b} \mid \mathbf{X})$

If s^2 is the estimate of σ^2 , a natural estimate of $\text{Var}(\mathbf{b} \mid \mathbf{X}) = \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$ is

$$\widehat{\text{Var}}(\mathbf{b} \mid \mathbf{X}) \equiv s^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}. \quad (1.3.4)$$

This is one of the statistics included in the computer printout of any OLS software package.

Questions for Review

- (Role of the no-multicollinearity assumption) In Propositions 1.1 and 1.2, where did we use Assumption 1.3 that $\text{rank}(\mathbf{X}) = K$? **Hint:** We need the no-multi-collinearity condition to make sure $\mathbf{X}'\mathbf{X}$ is invertible.
- (Example of a linear estimator) For the consumption function example in Example 1.1, propose a linear and unbiased estimator of β_2 that is different from the OLS estimator. **Hint:** How about $\widehat{\beta}_2 = (CON_2 - CON_1)/(YD_2 - YD_1)$? Is it linear in (CON_1, \dots, CON_n) ? Is it unbiased in the sense that $E(\widehat{\beta}_2 \mid YD_1, \dots, YD_n) = \beta_2$?
- (What Gauss-Markov doesn't mean) Under Assumption 1.1–1.4, does there exist a linear, but not necessarily unbiased, estimator of β which has a variance smaller than that of the OLS estimator? If so, how small can the variance be? **Hint:** If an estimator of β is a constant, then the estimator is trivially linear in \mathbf{y} .

4. (Gauss-Markov for Unconditional Variance)

(a) Prove: $\text{Var}(\hat{\boldsymbol{\beta}}) = \text{E}[\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X})] + \text{Var}[\text{E}(\hat{\boldsymbol{\beta}} | \mathbf{X})]$. **Hint:** By definition,

$$\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) \equiv \text{E}[(\hat{\boldsymbol{\beta}} - \text{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}))(\hat{\boldsymbol{\beta}} - \text{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}))' | \mathbf{X}]$$

and

$$\text{Var}[\text{E}(\hat{\boldsymbol{\beta}} | \mathbf{X})] \equiv \text{E}\{[\text{E}(\hat{\boldsymbol{\beta}} | \mathbf{X}) - \text{E}(\hat{\boldsymbol{\beta}})]^2\}.$$

Use the add-and-subtract strategy: take $\hat{\boldsymbol{\beta}} - \text{E}(\hat{\boldsymbol{\beta}} | \mathbf{X})$ and add-and-subtract $\text{E}(\hat{\boldsymbol{\beta}})$.

(b) Prove (1.3.3). **Hint:** If $\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X}) \geq \text{Var}(\mathbf{b} | \mathbf{X})$, then $\text{E}[\text{Var}(\hat{\boldsymbol{\beta}} | \mathbf{X})] \geq \text{E}[\text{Var}(\mathbf{b} | \mathbf{X})]$

5. Propose an unbiased estimator of σ^2 if you had data on $\boldsymbol{\varepsilon}$. **Hint:** How about $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/n$? Is it unbiased?

6. Prove part (d) of Proposition 1.1. **Hint:** By definition,

$$\text{Cov}(\mathbf{b}, \mathbf{e} | \mathbf{X}) \equiv \text{E}\left\{[\mathbf{b} - \text{E}(\mathbf{b} | \mathbf{X})][\mathbf{e} - \text{E}(\mathbf{e} | \mathbf{X})]' | \mathbf{X}\right\}.$$

Since $\text{E}(\mathbf{b} | \mathbf{X}) = \boldsymbol{\beta}$, we have: $\mathbf{b} - \text{E}(\mathbf{b} | \mathbf{X}) = \mathbf{A}\boldsymbol{\varepsilon}$ where \mathbf{A} here is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Use $\mathbf{M}\boldsymbol{\varepsilon} = \mathbf{e}$ (see Review Question 5 to Section 1.2) to show that $\mathbf{e} - \text{E}(\mathbf{e} | \mathbf{X}) = \mathbf{M}\boldsymbol{\varepsilon}$. $\text{E}(\mathbf{A}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M} | \mathbf{X}) = \mathbf{A}\text{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X})\mathbf{M}$ since both \mathbf{A} and \mathbf{M} are functions of \mathbf{X} . Finally use $\mathbf{M}\mathbf{X} = \mathbf{0}$ (see (1.2.11)).

7. Prove (1.2.21). **Hint:** Since \mathbf{P} is positive semi-definite, its diagonal elements are nonnegative. Note that $\sum_{i=1}^n p_i = \text{trace}(\mathbf{P})$.

1.4 Hypothesis Testing under Normality

Very often, the economic theory that motivated the regression equation also specifies the values that the regression coefficients should take. Suppose that the underlying theory implies the restriction that β_2 equals 1. Although Proposition 1.1 guarantees that, on average, b_2 (the OLS estimate of β_2) equals 1 if the restriction is true, b_2 may not be exactly equal to 1 for a particular sample at hand. Obviously, we cannot conclude that the restriction is false just because the estimate b_2 differs from 1. In order for us to decide whether the sampling error $b_2 - 1$ is “too large” for

the restriction to be true, we need to construct from the sampling error some test statistic whose probability distribution is known given the truth of the hypothesis. It might appear that doing so requires one to specify the joint distribution of $(\mathbf{X}, \varepsilon)$ because, as clear from (1.2.14), the sampling error is a function of $(\mathbf{X}, \varepsilon)$. A surprising fact about the theory of hypothesis testing to be presented in this section is that the distribution can be derived without specifying the joint distribution when the conditional distribution of ε conditional on \mathbf{X} is normal; there is no need to specify the distribution of \mathbf{X} .

In the language of hypothesis testing, the restriction to be tested (such as “ $\beta_2 = 1$ ”) is called the **null hypothesis** (or simply the **null**). It is a restriction on the **maintained hypothesis**, a set of assumptions which, combined with the null, produces some test statistic with a known distribution. For the present case of testing hypothesis about regression coefficients, only the normality assumption about the conditional distribution of ε needs to be added to the classical regression model (which is Assumptions 1.1–1.4) to form the maintained hypothesis (as just noted, there is no need to specify the joint distribution of $(\mathbf{X}, \varepsilon)$). Sometimes the maintained hypothesis is (somewhat loosely) referred to as “the model”. We say that the model is **correctly specified** if the maintained hypothesis is true. Although too large a value of the test statistic is interpreted as a failure of the null, the interpretation is valid only as long as the model is correctly specified. It is possible that the test statistic does not have the supposed distribution when the null is true but the model is false.

Normally Distributed Error Terms

In many applications, the error term consists of many miscellaneous factors not captured by the regressors. The Central Limit Theorem suggests that the error term has a normal distribution. In other applications, the error term is due to errors in measuring the dependent variable. It is known that very often measurement errors are normally distributed (in fact, the normal distribution was originally developed for measurement errors). It is therefore worth entertaining the normality assumption:

Assumption 1.5 (normality of the error term): *The distribution of ε conditional on \mathbf{X} is jointly normal.*

We recall from probability theory that the normal distribution has several convenient features.

- The distribution depends only on the mean and the variance. Thus, once the mean and the variance are known, you can write down the density function.

If the distribution conditional on \mathbf{X} is normal, the mean and the variance can depend on \mathbf{X} . It follows that, if the distribution conditional on \mathbf{X} is normal and if neither the conditional mean nor the conditional variance depends on \mathbf{X} , then the marginal (i.e., unconditional) distribution is the same normal distribution.

- In general, if two random variables are independent, then they are uncorrelated but the converse is not true. If two random variables are joint normal, the converse is also true so that independence and a lack of correlation are equivalent. This carries over to conditional distributions: if two random variables are joint normal and uncorrelated conditional on \mathbf{X} , then they are independent conditional on \mathbf{X} .
- A linear function of random variables that are jointly normally distributed is itself normally distributed. This also carries over to conditional distributions. If the distribution of $\boldsymbol{\varepsilon}$ conditional on \mathbf{X} is normal, then $\mathbf{A}\boldsymbol{\varepsilon}$, where the elements of matrix \mathbf{A} are functions of \mathbf{X} , is normal conditional on \mathbf{X} .

It is thanks to these features of normality that Assumption 1.5 delivers the following properties to be exploited in the derivation of test statistics.

- The mean and the variance of the distribution of $\boldsymbol{\varepsilon}$ conditional on \mathbf{X} are already specified in Assumptions 1.2 and 1.4. Therefore, Assumption 1.5 together with Assumptions 1.2 and 1.4 implies that the distribution of $\boldsymbol{\varepsilon}$ conditional on \mathbf{X} is $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$:

$$\boldsymbol{\varepsilon} \mid \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n). \quad (1.4.1)$$

Thus, the distribution of $\boldsymbol{\varepsilon}$ conditional on \mathbf{X} does not depend on \mathbf{X} . It then follows that $\boldsymbol{\varepsilon}$ and \mathbf{X} are *independent*. Therefore, in particular, the marginal or unconditional distribution of $\boldsymbol{\varepsilon}$ is $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$.

- We know from (1.2.14) that the sampling error $\mathbf{b} - \boldsymbol{\beta}$ is linear in $\boldsymbol{\varepsilon}$ given \mathbf{X} . Since $\boldsymbol{\varepsilon}$ is normal given \mathbf{X} , so is the sampling error. Its mean and variance are given by parts (a) and (b) of Proposition 1.1. Thus, under Assumptions 1.1–1.5,

$$(\mathbf{b} - \boldsymbol{\beta}) \mid \mathbf{X} \sim N(\mathbf{0}, \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}). \quad (1.4.2)$$

Testing Hypotheses about Individual Regression Coefficients

The type of hypotheses we first consider is about the k -th coefficient

$$H_0: \beta_k = \bar{\beta}_k.$$

Here, $\bar{\beta}_k$ is some known value specified by the null hypothesis. We wish to test this null against the alternative hypothesis $H_1: \beta_k \neq \bar{\beta}_k$, at a significance level of α . Looking at the k -th component of (1.4.2) and imposing the restriction of the null, we obtain

$$(b_k - \bar{\beta}_k) \mid \mathbf{X} \sim N\left(0, \sigma^2 \cdot ((\mathbf{X}'\mathbf{X})^{-1})_{kk}\right),$$

where $((\mathbf{X}'\mathbf{X})^{-1})_{kk}$ is the (k, k) element of $(\mathbf{X}'\mathbf{X})^{-1}$. So if we define the ratio z_k by dividing $b_k - \bar{\beta}_k$ by its standard deviation:

$$z_k \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{\sigma^2 \cdot ((\mathbf{X}'\mathbf{X})^{-1})_{kk}}}, \quad (1.4.3)$$

then the distribution of z_k is $N(0, 1)$ (the standard normal distribution).

Suppose for a second that σ^2 is known. Then the statistic z_k has some desirable properties as a test statistic. First, its value can be calculated from the sample. Second, its distribution conditional on \mathbf{X} does not depend on \mathbf{X} (which should not be confused with that fact that the *value* of z_k depends on \mathbf{X}). So z_k and \mathbf{X} are independently distributed and, regardless of the value of \mathbf{X} , the distribution of z_k is the same as its unconditional distribution. This is convenient because different samples differ not only in \mathbf{y} but also in \mathbf{X} . Third, the distribution is known. In particular, it does not depend on unknown parameters (such as β). (If the distribution of a statistic depends on unknown parameters, those parameters are called **nuisance parameters**.) Using this statistic, we can determine whether or not the sampling error $b_k - \bar{\beta}_k$ is “too large”: it is too large if the test statistic takes on a value that is surprising for a realization from the distribution.

If we don't know the true value of σ^2 , a natural idea is to replace the nuisance parameter σ^2 by its OLS estimate s^2 . The statistic after the substitution of s^2 for σ^2 is called the ***t*-ratio** or the ***t*-value**. The denominator of this statistic is called the **standard error** of the OLS estimate of β_k and sometimes written as $SE(b_k)$:

$$SE(b_k) \equiv \sqrt{s^2 \cdot ((\mathbf{X}'\mathbf{X})^{-1})_{kk}} = \sqrt{(k, k) \text{ element of } \widehat{\text{Var}}(\mathbf{b} \mid \mathbf{X}) \text{ in (1.3.4)}}. \quad (1.4.4)$$

Since s^2 , being a function of the sample, is a random variable, this substitution changes the distribution of the statistic, but fortunately the changed distribution, too, is known and depends on neither nuisance parameters nor \mathbf{X} .

Proposition 1.3 (distribution of the *t*-ratio): *Suppose Assumptions 1.1–1.5 hold. Under the null hypothesis $H_0: \beta_k = \bar{\beta}_k$, the *t*-ratio defined as*

$$t_k \equiv \frac{b_k - \bar{\beta}_k}{SE(b_k)} \equiv \frac{b_k - \bar{\beta}_k}{\sqrt{s^2 \cdot ((\mathbf{X}'\mathbf{X})^{-1})_{kk}}} \quad (1.4.5)$$

*is distributed as $t(n - K)$ (the *t* distribution with $n - K$ degrees of freedom).*

PROOF. We can write

$$\begin{aligned} t_k &= \frac{b_k - \beta_k}{\sqrt{\sigma^2 \cdot ((\mathbf{X}'\mathbf{X})^{-1})_{kk}}} \cdot \sqrt{\frac{\sigma^2}{s^2}} = \frac{z_k}{\sqrt{s^2/\sigma^2}} \\ &= \frac{z_k}{\sqrt{\frac{\mathbf{e}'\mathbf{e}/(n-K)}{\sigma^2}}} = \frac{z_k}{\sqrt{\frac{q}{n-K}}}, \end{aligned}$$

where $q \equiv \mathbf{e}'\mathbf{e}/\sigma^2$ to reflect the substitution of s^2 for σ^2 . We have already shown that z_k is $N(0, 1)$. We will show:

- (1) $q \mid \mathbf{X} \sim \chi^2(n - K)$,
- (2) two random variables z_k and q are independent conditional on \mathbf{X} .

Then, by the definition of the t distribution, the ratio of z_k to $\sqrt{q/(n - K)}$ is distributed as t with $n - K$ degrees of freedom,¹⁵ and we are done.

- (1) Since $\mathbf{e}'\mathbf{e} = \boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}$ from (1.2.12), we have

$$q = \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}}{\sigma}.$$

The middle matrix \mathbf{M} , being the annihilator, is idempotent. Also, $\boldsymbol{\varepsilon}/\sigma \mid \mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_n)$ by (1.4.1). Therefore, this quadratic form is distributed as χ^2 with degrees of freedom equal to $\text{rank}(\mathbf{M})$.¹⁶ But $\text{rank}(\mathbf{M}) = \text{trace}(\mathbf{M})$, because \mathbf{M} is idempotent.¹⁷ We have already shown in the proof of Proposition 1.2 that $\text{trace}(\mathbf{M}) = n - K$. So $q \mid \mathbf{X} \sim \chi^2(n - K)$.

- (2) Both \mathbf{b} and \mathbf{e} are linear functions of $\boldsymbol{\varepsilon}$ (by (1.2.14) and the fact that $\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$), so they are jointly normal conditional on \mathbf{X} . Also, they are uncorrelated conditional on \mathbf{X} (see part (d) of Proposition 1.1). So \mathbf{b} and \mathbf{e} are independently distributed conditional on \mathbf{X} . But z_k is a function of \mathbf{b} and q is a function of \mathbf{e} . So z_k and q are independently distributed conditional on \mathbf{X} .¹⁸ ■

Decision Rule for the t -Test

The test of the null hypothesis based on the t -ratio is called the t -test and proceeds as follows.

¹⁵Fact: if $x \sim N(0, 1)$, $y \sim \chi^2(m)$ and if x and y are independent, then the ratio $x/\sqrt{y/m}$ has the t distribution with m degrees of freedom.

¹⁶Fact: If $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_n)$ and \mathbf{A} is idempotent, then $\mathbf{x}'\mathbf{A}\mathbf{x}$ has a chi-squared distribution with degrees of freedom equal to the rank of \mathbf{A} .

¹⁷Fact: If \mathbf{A} is idempotent, then $\text{rank}(\mathbf{A}) = \text{trace}(\mathbf{A})$.

¹⁸Fact: If \mathbf{x} and \mathbf{y} are independently distributed, then so are $f(\mathbf{x})$ and $g(\mathbf{y})$.

Step 1: Given the hypothesized value, $\bar{\beta}_k$, of β_k , form the t -ratio as in (1.4.5). Too large a deviation of t_k from 0 is a sign of the failure of the null hypothesis. The next step specifies how large is “too large”.

Step 2: Go to the t -table (most statistics and econometrics textbooks include the t -table) and look up the entry for $n - K$ degrees of freedom. Find the **critical value**, $t_{\alpha/2}(n - K)$, such that the area in the t distribution to the right of $t_{\alpha/2}(n - K)$ is $\alpha/2$, as illustrated in Figure 1.3. (If $n - K = 30$ and $\alpha = 5\%$, for example, $t_{\alpha/2}(n - K) = 2.042$.) Then, since the t distribution is symmetric around 0,

$$\text{Prob}(-t_{\alpha/2}(n - K) < t < t_{\alpha/2}(n - K)) = 1 - \alpha.$$

Step 3: Accept H_0 if $-t_{\alpha/2}(n - K) < t_k < t_{\alpha/2}(n - K)$ (that is, if $|t_k| < t_{\alpha/2}(n - K)$), where t_k is the t -ratio from *Step 1*. Reject H_0 otherwise. Since $t_k \sim t(n - K)$ under H_0 , the probability of rejecting H_0 when H_0 is true is α . So the size (significance level) of the test is indeed α .

A convenient feature of the t test is that the critical value does not depend on \mathbf{X} ; there is no need to calculate critical values for each sample.

Confidence Interval

Step 3 can also be stated in terms of b_k and $SE(b_k)$. Since t_k is as in (1.4.5), you accept H_0 whenever

$$-t_{\alpha/2}(n - K) < \frac{b_k - \bar{\beta}_k}{SE(b_k)} < t_{\alpha/2}(n - K)$$

or

$$b_k - SE(b_k) \cdot t_{\alpha/2}(n - K) < \bar{\beta}_k < b_k + SE(b_k) \cdot t_{\alpha/2}(n - K).$$

Therefore, we accept if and only if the hypothesized value $\bar{\beta}_k$ falls in the interval:

$$[b_k - SE(b_k) \cdot t_{\alpha/2}(n - K), b_k + SE(b_k) \cdot t_{\alpha/2}(n - K)]. \quad (1.4.6)$$

This interval is called the **level $1 - \alpha$ confidence interval**. It is narrower the smaller the standard error. Thus, the smallness of the standard error is a measure of the estimator’s precision.

p -Value

The decision rule of the t -test can also be stated using the **p -value**.

Step 1: Same as above.

Step 2: Rather than finding the critical value $t_{\alpha/2}(n - K)$, calculate

$$p = \text{Prob}(t > |t_k|) \times 2.$$

Since the t distribution is symmetric around 0, $\text{Prob}(t > |t_k|) = \text{Prob}(t < -|t_k|)$, so

$$\text{Prob}(-|t_k| < t < |t_k|) = 1 - p. \quad (1.4.7)$$

Step 3: Accept H_0 if $p > \alpha$. Reject otherwise.

To see the equivalence of the two decision rules, one based on the critical values such as $t_{\alpha/2}(n - K)$ and the other based on the p -value, refer to Figure 1.3. If $\text{Prob}(t > |t_k|)$ is greater than $\alpha/2$ (as in the figure), that is, if the p -value is more than α , then $|t_k|$ must be to the left of $t_{\alpha/2}(n - K)$. This means from *Step 3* that the null hypothesis is not rejected. Thus, when p is small, the t -ratio is surprisingly large for a random variable from the t distribution. The smaller the p , the stronger the rejection.

Examples of the t -test can be found in Section 1.7.

Linear Hypotheses

The null hypothesis we wish to test may not be a restriction about individual regression coefficients of the maintained hypothesis; it is often about linear combinations of them written as a system of linear equations:

$$H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}, \quad (1.4.8)$$

where values of \mathbf{R} and \mathbf{r} are known and specified by the hypothesis. We denote the number of equations, which is the dimension of \mathbf{r} , by $\#\mathbf{r}$. So \mathbf{R} is $\#\mathbf{r} \times K$. These $\#\mathbf{r}$ equations are restrictions on the coefficients in the maintained hypothesis. It is called a linear hypothesis because each equation is linear. To make sure that there are no redundant equations and that the equations are consistent with each other, we require that $\text{rank}(\mathbf{R}) = \#\mathbf{r}$ (i.e., \mathbf{R} is of full *row* rank with its rank equaling the number of rows). But don't be too conscious about the rank condition; in specific applications, it is very easy to spot a failure of the rank condition if there is one.

Example 1.5 (continuation of Example 1.2): Consider the wage equation of Example 1.2 where $K = 4$. We might wish to test the hypothesis that education and tenure have equal impact on the wage rate and that there is no experience effect. The hypothesis is two equations (so $\#\mathbf{r} = 2$):

$$\beta_2 = \beta_3 \quad \text{and} \quad \beta_4 = 0.$$

This can be cast in the format $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ if \mathbf{R} and \mathbf{r} are defined as

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Because the two rows of this \mathbf{R} are linearly independent, the rank condition is satisfied.

But suppose we require additionally that

$$\beta_2 - \beta_3 = \beta_4.$$

This is redundant because it holds whenever the first two equations do. With these three equations, $\#\mathbf{r} = 3$ and

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & -1 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Since the third row of \mathbf{R} is the difference between the first two, \mathbf{R} is not of full row rank. The consequence of adding redundant equations is that \mathbf{R} no longer meets the full row rank condition.

As an example of inconsistent equations, consider adding to the first two equations the third equation $\beta_4 = 0.5$. Evidently, β_4 cannot be 0 and 0.5 at the same time. The hypothesis is inconsistent because there is no $\boldsymbol{\beta}$ that satisfies the three equations simultaneously. If we nevertheless included this equation, then \mathbf{R} and \mathbf{r} would become

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{r} = \begin{bmatrix} 0 \\ 0 \\ 0.5 \end{bmatrix}.$$

Again, the full row rank condition is not satisfied because the rank of \mathbf{R} is 2 while $\#\mathbf{r} = 3$.

The F -Test

In order to test linear hypotheses, we look for a test statistic that has a known distribution under the null hypothesis.

Proposition 1.4 (distribution of the F -ratio): Suppose Assumptions 1.1–1.5 hold. Under the null hypothesis $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, where \mathbf{R} is $\#\mathbf{r} \times K$ with $\text{rank}(\mathbf{R}) = \#\mathbf{r}$, the F -ratio defined as

$$\begin{aligned} F &\equiv \frac{(\mathbf{R}\mathbf{b} - \mathbf{r})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} (\mathbf{R}\mathbf{b} - \mathbf{r}) / \#\mathbf{r}}{s^2} \\ &= (\mathbf{R}\mathbf{b} - \mathbf{r})' [\widehat{\mathbf{R}\text{Var}(\mathbf{b} \mid \mathbf{X})\mathbf{R}'}]^{-1} (\mathbf{R}\mathbf{b} - \mathbf{r}) / \#\mathbf{r} \quad (\text{by (1.3.4)}) \end{aligned} \tag{1.4.9}$$

is distributed as $F(\#r, n - K)$ (the F distribution with $\#r$ and $n - K$ degrees of freedom).

As in Proposition 1.3, it suffices to show that the distribution conditional on \mathbf{X} is $F(\#r, n - K)$; because the F distribution doesn't depend on \mathbf{X} , it is also the unconditional distribution of the statistic.

PROOF. Since $s^2 = \mathbf{e}'\mathbf{e}/(n - K)$, we can write

$$F = \frac{w/\#r}{q/(n - K)}$$

where

$$w \equiv (\mathbf{R}\mathbf{b} - \mathbf{r})'[\sigma^2 \cdot \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) \quad \text{and} \quad q \equiv \frac{\mathbf{e}'\mathbf{e}}{\sigma^2}.$$

We need to show:

- (1) $w \mid \mathbf{X} \sim \chi^2(\#r)$,
- (2) $q \mid \mathbf{X} \sim \chi^2(n - K)$ (this part you were asked to show for Proposition 1.3),
- (3) w and q are independently distributed conditional on \mathbf{X} .

Then, by the definition of the F distribution, the F -ratio $\sim F(\#r, n - K)$.

- (1) Let $\mathbf{v} \equiv \mathbf{R}\mathbf{b} - \mathbf{r}$. Under H_0 , $\mathbf{R}\mathbf{b} - \mathbf{r} = \mathbf{R}(\mathbf{b} - \boldsymbol{\beta})$. So by (1.4.2), conditional on \mathbf{X} , \mathbf{v} is normal with mean $\mathbf{0}$, and its variance is given by

$$\text{Var}(\mathbf{v} \mid \mathbf{X}) = \text{Var}(\mathbf{R}(\mathbf{b} - \boldsymbol{\beta}) \mid \mathbf{X}) = \mathbf{R} \text{Var}(\mathbf{b} - \boldsymbol{\beta} \mid \mathbf{X})\mathbf{R}' = \sigma^2 \cdot \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}',$$

which is none other than the inverse of the middle matrix in the quadratic form for w . Hence w can be written as $\mathbf{v}' \text{Var}(\mathbf{v} \mid \mathbf{X})^{-1}\mathbf{v}$. Since \mathbf{R} is of full row rank and $\mathbf{X}'\mathbf{X}$ is nonsingular, $\sigma^2 \cdot \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ is nonsingular (why? Showing this is a review question). Therefore, by the definition of the χ^2 distribution, $w \mid \mathbf{X} \sim \chi^2(\#r)$.¹⁹

- (3) w is a function of \mathbf{b} and q is a function of \mathbf{e} . But \mathbf{b} and \mathbf{e} are independently distributed conditional on \mathbf{X} , as shown in part (2) of the proof of Proposition 1.3. So w and q are independently distributed conditional on \mathbf{X} . ■

If the null hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ is true, we expect $\mathbf{R}\mathbf{b} - \mathbf{r}$ to be small, so large values of F should be taken as evidence for a failure of the null. This means that we look at only the upper tail of the distribution in the F -statistic. The decision rule of the F -test at the significance level of α is as follows.

¹⁹Fact: Let \mathbf{x} be an m dimensional random vector. If $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ nonsingular, then $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \sim \chi^2(m)$.

Step 1: Calculate the F -ratio by the formula (1.4.9).

Step 2: Go to the table of F distribution and look up the entry for $\#r$ (the numerator degrees of freedom) and $n - K$ (the denominator degrees of freedom). Find the critical value $F_\alpha(\#r, n - K)$ that leaves α for the upper tail of the F distribution, as illustrated in Figure 1.4. For example, when $\#r = 3$, $n - K = 30$, and $\alpha = 5\%$, the critical value $F_{.05}(3, 30)$ is 2.92.

Step 3: Accept the null if the F -ratio from *Step 1* is less than $F_\alpha(\#r, n - K)$. Reject otherwise.

This decision rule can also be described in terms of the p -value.

Step 1: Same as above.

Step 2: Calculate

$$p = \text{area of the upper tail of the } F \text{ distribution to the right of the } F\text{-ratio.}$$

Step 3: Accept the null if $p > \alpha$; reject otherwise.

Thus a *small* p -value is a signal of the failure of the null.

A More Convenient Expression for F

The above derivation of the F -ratio is by the **Wald principle**, because it is based on the unrestricted estimator which is not constrained to satisfy the restrictions of the null hypothesis. Calculating the F -ratio by the formula (1.4.9) requires matrix inversion and multiplication. Fortunately, there is a convenient alternative formula involving two different sum of squared residuals: one is SSR , the minimized sum of squared residuals obtained from (1.2.1) now denoted as SSR_U , and the other is the restricted sum of squared residuals, denoted SSR_R , obtained from

$$\min_{\tilde{\beta}} SSR(\tilde{\beta}) \quad \text{s.t.} \quad \mathbf{R}\tilde{\beta} = \mathbf{r}. \quad (1.4.10)$$

Finding the $\tilde{\beta}$ that achieves this constrained minimization is called the **restricted regression** or **restricted least squares**. It is left as an analytical exercise to show that the F -ratio equals:

$$F = \frac{(SSR_R - SSR_U)/\#r}{SSR_U/(n - K)}, \quad (1.4.11)$$

which is the difference in the objective function deflated by the estimate of the error variance. This derivation of the F -ratio is analogous to how the likelihood-ratio statistic is derived in maximum likelihood estimation as the difference in

log likelihood with and without the imposition of the null hypothesis. For this reason, this second derivation of the F -ratio is said to be by the **Likelihood-Ratio principle**. There is a closed-form expression for the restricted least squares estimator of β . Deriving the expression is left as an analytical exercise. The computation of restricted least squares will be explained in the context of the empirical example in Section 1.7.

t versus F

Because hypotheses about individual coefficients are linear hypotheses, the t -test of $H_0: \beta_k = \bar{\beta}_k$ is a special case of the F -test. To see this, note that the hypothesis can be written as $\mathbf{R}\beta = \mathbf{r}$ with

$$\underset{(1 \times K)}{\mathbf{R}} = \begin{bmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{bmatrix}, \quad \mathbf{r} = \bar{\beta}_k.$$

(k)

So by (1.4.9) the F -ratio is

$$F = (b_k - \bar{\beta}_k) [s^2 \cdot (k, k) \text{ element of } (\mathbf{X}'\mathbf{X})^{-1}]^{-1} (b_k - \bar{\beta}_k),$$

which is the square of the t -ratio in (1.4.5). Since a random variable distributed as $F(1, n - K)$ is the square of a random variable distributed as $t(n - K)$, the t - and F -tests give the same test result.

Sometimes, the null is that a set of individual regression coefficients equal certain values. For example, assume $K = 2$ and consider

$$H_0: \beta_1 = 1 \quad \text{and} \quad \beta_2 = 0.$$

This can be written as a linear hypothesis $\mathbf{R}\beta = \mathbf{r}$ for $\mathbf{R} = \mathbf{I}_2$ and $\mathbf{r} = (1, 0)'$. So the F -test can be used. It is tempting, however, to conduct the t -test separately for each individual coefficient of the hypothesis. We might accept H_0 if both restrictions $\beta_1 = 1$ and $\beta_2 = 0$ pass the t -test. This amounts to using the confidence region of

$$\left\{ (\beta_1, \beta_2) \mid \begin{aligned} b_1 - SE(b_1) \cdot t_{\alpha/2}(n - K) < \beta_1 < b_1 + SE(b_1) \cdot t_{\alpha/2}(n - K), \\ b_2 - SE(b_2) \cdot t_{\alpha/2}(n - K) < \beta_2 < b_2 + SE(b_2) \cdot t_{\alpha/2}(n - K) \end{aligned} \right\},$$

which is a rectangular region in the (β_1, β_2) plane, as illustrated in Figure 1.5. If $(1, 0)$, the point in the (β_1, β_2) plane specified by the null falls in this region, one would accept the null. On the other hand, the confidence region for the F -test is

$$\left\{ (\beta_1, \beta_2) \mid (b_1 - \beta_1, b_2 - \beta_2) \left(\widehat{\text{Var}}(\mathbf{b} \mid \mathbf{X}) \right)^{-1} \begin{bmatrix} b_1 - \beta_1 \\ b_2 - \beta_2 \end{bmatrix} < 2F_{\alpha}(\#\mathbf{r}, n - K) \right\},$$

Since $\widehat{\text{Var}}(\mathbf{b} \mid \mathbf{X})$ is positive definite, the F -test acceptance region is an ellipse in the (β_1, β_2) plane. The two confidence regions look typically like Figure 1.5.

The F -test should be preferred to the test using two t -ratios for two reasons. First, if the size (significance level) in each of the two t -tests is α , then the overall size (the probability that $(1, 0)$ is outside the rectangular region) is not α . Second, as will be noted in the next section (see (1.5.19)), the F -test is a likelihood ratio test and likelihood-ratio tests have certain desirable properties. So even if the significance level in each t -test is controlled so that the overall size is α , the test is less desirable than the F -test.²⁰

An Example of a Test Statistic Whose Distribution Depends on \mathbf{X}

To place the discussion of this section in a proper perspective, it may be useful to note that there are some statistics whose conditional distribution depends on \mathbf{X} . Consider the celebrated **Durbin-Watson statistic**:

$$\frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}.$$

The conditional distribution, and hence the critical values, of this statistic depend on \mathbf{X} , but J. Durbin and G. S. Watson have shown that the critical values fall between two bounds (which depends on the sample size, the number of regressors, and whether the regressor includes a constant). Therefore, the critical values for the unconditional distribution, too, fall between these bounds.

The statistic is designed for testing whether there is no serial correlation in the error term. Thus the null hypothesis is Assumption 1.4, while the maintained hypothesis is the other assumptions of the classical regression model (including the strict exogeneity assumption) and the normality assumption. But, as emphasized in Section 1.1, the strict exogeneity assumption is not satisfied in time-series models typically encountered in econometrics, and serial correlation is an issue that arises only in time-series models. Thus the Durbin-Watson statistic is not useful in econometrics. More useful tests for serial correlation, which are all based on large-sample theory, will be covered in the next chapter.

²⁰For more details on the relationship between the t -test and the F -tests, see Scheffe (1959, p. 46).

Questions for Review

- (Conditional vs. unconditional distribution) Do we know from Assumptions 1.1–1.5 that the marginal (unconditional) distribution of \mathbf{b} is normal? [Answer: No.] Are the statistics z_k (see (1.4.3)), t_k , and F distributed independently of \mathbf{X} ? [Answer: Yes, because their distributions conditional on \mathbf{X} don't depend on \mathbf{X} .]
- (Computation of test statistics) Verify that $SE(b_k)$ as well as \mathbf{b} , SSR , s^2 , and R^2 can be calculated from the following sample averages: $\mathbf{S}_{\mathbf{xx}}$, $\mathbf{s}_{\mathbf{xy}}$, $\mathbf{y}'\mathbf{y}/n$, and \bar{y} .
- For the formula (1.4.9) for the F to be well-defined, the matrix $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ must be nonsingular. Prove the stronger result that the matrix is positive definite. **Hint:** $\mathbf{X}'\mathbf{X}$ is positive definite. The inverse of a positive definite matrix is positive definite. Since $\mathbf{R}(\#r \times K)$ is of full row rank, for any non-zero $\#r$ dimensional vector \mathbf{z} , $\mathbf{R}'\mathbf{z} \neq \mathbf{0}$.
- (One-tailed t -test) The t -test described in the text is the **two-tailed t -test** because the significance α is equally distributed between both tails of the t distribution. Suppose the alternative is one-sided and written as $H_1: \beta_k > \bar{\beta}_k$. Consider the following modification of the decision rule of the t -test.

Step 1: Same as above.

Step 2: Find the critical value t_α such that the area in the t distribution to the right of t_α is α . Note the difference from the two-tailed test: the left tail is ignored and the area of α is assigned to the upper tail only.

Step 3: Accept if $t_k < t_\alpha$; reject otherwise.

Show that the size (significance level) of this **one-tailed t -test** is α .

- (Relation between $F(1, n-K)$ and $t(n-K)$) Look up the t and F distribution tables to verify that $F_\alpha(1, n-K) = (t_{\alpha/2}(n-K))^2$ for degrees of freedom and significance levels of your choice.
- (t vs. F) “It is nonsense to test a hypothesis consisting of a large number of equality restrictions, because the t -test will most likely reject at least some of the restrictions.” Criticize this statement.

7. (Variance of s^2) Show that, under Assumptions 1.1–1.5,

$$\text{Var}(s^2 \mid \mathbf{X}) = \frac{2\sigma^4}{n - K}.$$

Hint: If a random variable is distributed as $\chi^2(m)$, then its mean is m and variance $2m$.

1.5 Relation to Maximum Likelihood

Having specified the distribution of the error vector $\boldsymbol{\varepsilon}$, we can use the **maximum likelihood (ML) principle** to estimate the model parameters $(\boldsymbol{\beta}, \sigma^2)$.²¹ In this section, we will show that \mathbf{b} , the OLS estimator of $\boldsymbol{\beta}$, is also the ML estimator and the OLS estimator of σ^2 differs only slightly from the ML counterpart, when the error is normally distributed. We will also show that \mathbf{b} achieves the **Cramer-Rao lower bound**.

The Maximum Likelihood Principle

As you might recall from elementary statistics, the basic idea of the ML principle is to choose the parameter estimates to maximize the probability of obtaining the observed sample. To be more precise, we assume that the probability density of the sample (\mathbf{y}, \mathbf{X}) is a member of a family of functions indexed by a finite-dimensional parameter vector $\tilde{\boldsymbol{\zeta}}$: $f(\mathbf{y}, \mathbf{X}; \tilde{\boldsymbol{\zeta}})$. (This is described as **parameterizing** the density function.) This function, viewed as a function of the hypothetical parameter vector $\tilde{\boldsymbol{\zeta}}$, is called the **likelihood function**. At the true parameter vector $\boldsymbol{\zeta}$, the density of (\mathbf{y}, \mathbf{X}) is $f(\mathbf{y}, \mathbf{X}; \boldsymbol{\zeta})$. The ML estimate of the true parameter vector $\boldsymbol{\zeta}$ is the $\tilde{\boldsymbol{\zeta}}$ that maximizes the likelihood function given the data (\mathbf{y}, \mathbf{X}) .

Conditional vs. Unconditional Likelihood

Since a (joint) density is the product of a marginal density and a conditional density, the density of (\mathbf{y}, \mathbf{X}) can be written as

$$f(\mathbf{y}, \mathbf{X}; \boldsymbol{\zeta}) = f(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\theta}) \cdot f(\mathbf{X}; \boldsymbol{\psi}), \quad (1.5.1)$$

where $\boldsymbol{\theta}$ is the subset of the parameter vector $\boldsymbol{\zeta}$ that determine the conditional density function and $\boldsymbol{\psi}$ is the subset determining the marginal density function.

²¹For a fuller treatment of maximum likelihood, see Chapter 7.

The parameter vector of interest is $\boldsymbol{\theta}$; for the linear regression model with normal errors, $\boldsymbol{\theta} = (\boldsymbol{\beta}', \sigma^2)'$ and $f(\mathbf{y} | \mathbf{X}; \boldsymbol{\theta})$ is given by (1.5.4) below.

Let $\tilde{\boldsymbol{\zeta}} \equiv (\tilde{\boldsymbol{\theta}}', \tilde{\boldsymbol{\psi}}')'$ be a hypothetical value of $\boldsymbol{\zeta} = (\boldsymbol{\theta}', \boldsymbol{\psi}')'$. Then the (unconditional or joint) likelihood function is

$$f(\mathbf{y}, \mathbf{X}; \tilde{\boldsymbol{\zeta}}) = f(\mathbf{y} | \mathbf{X}; \tilde{\boldsymbol{\theta}}) \cdot f(\mathbf{X}; \tilde{\boldsymbol{\psi}}). \quad (1.5.2)$$

If we knew the parametric form of $f(\mathbf{X}; \tilde{\boldsymbol{\psi}})$, then we could maximize this joint likelihood function over the entire hypothetical parameter vector $\tilde{\boldsymbol{\zeta}}$ and the ML estimate of $\boldsymbol{\theta}$ would be the elements of the ML estimate of $\boldsymbol{\zeta}$. We cannot do this for the classical regression model because the model doesn't specify $f(\mathbf{X}; \tilde{\boldsymbol{\psi}})$. However, if there is no functional relationship between $\tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\psi}}$ (such as a subset of $\tilde{\boldsymbol{\psi}}$ being a function of $\tilde{\boldsymbol{\theta}}$), then maximizing (1.5.2) with respect to $\tilde{\boldsymbol{\zeta}}$ is achieved by separately maximizing $f(\mathbf{y} | \mathbf{X}; \tilde{\boldsymbol{\theta}})$ with respect to $\tilde{\boldsymbol{\theta}}$ and maximizing $f(\mathbf{X}; \tilde{\boldsymbol{\psi}})$ with respect to $\tilde{\boldsymbol{\psi}}$. Thus the ML estimate of $\boldsymbol{\theta}$ also maximizes the **conditional likelihood** $f(\mathbf{y} | \mathbf{X}; \tilde{\boldsymbol{\theta}})$.

The Log Likelihood for the Regression Model

As already observed, Assumptions 1.5 (the normality assumption) together with Assumptions 1.2 and 1.4 implies that the distribution of $\boldsymbol{\varepsilon}$ conditional on \mathbf{X} is $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ (see (1.4.1)). But since $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ by Assumption 1.1, we have

$$\mathbf{y} | \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n). \quad (1.5.3)$$

Thus the conditional density of \mathbf{y} given \mathbf{X} is²²

$$f(\mathbf{y} | \mathbf{X}) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]. \quad (1.5.4)$$

Replacing the true parameters $(\boldsymbol{\beta}, \sigma^2)$ by their hypothetical values $(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)$ and taking logs, we obtain the **log likelihood function**:

$$\log L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}). \quad (1.5.5)$$

Since the log transformation is a monotone transformation, the ML estimator of $(\boldsymbol{\beta}, \sigma^2)$ is the $(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)$ that maximizes this log likelihood.

²²Recall from basic probability theory that the density function for an n -variate normal distribution with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ is

$$(2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right].$$

To derive (1.5.4), just set $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$.

ML via Concentrated Likelihood

It is instructive to maximize the log likelihood in two stages. First, maximize over $\tilde{\boldsymbol{\beta}}$ for any given $\tilde{\sigma}^2$. The $\tilde{\boldsymbol{\beta}}$ that maximizes the objective function could (but does not, in the present case of Assumptions 1.1–1.5) depend on $\tilde{\sigma}^2$. Second, maximize over $\tilde{\sigma}^2$ taking into account that the $\tilde{\boldsymbol{\beta}}$ obtained in the first stage could depend on $\tilde{\sigma}^2$. The log likelihood function in which $\tilde{\boldsymbol{\beta}}$ is constrained to be the value from the first stage is called the **concentrated log likelihood function** (concentrated with respect to $\tilde{\boldsymbol{\beta}}$). For the normal log likelihood (1.5.5), the first stage amounts to minimizing the sum of squares $(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$. The $\tilde{\boldsymbol{\beta}}$ that does it is none other than the OLS estimator \mathbf{b} and the minimized sum of squares is $\mathbf{e}'\mathbf{e}$. Thus the concentrated log likelihood is

$$\text{concentrated log likelihood} = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2} \mathbf{e}'\mathbf{e}. \quad (1.5.6)$$

This is a function of $\tilde{\sigma}^2$ alone, and the $\tilde{\sigma}^2$ that maximizes the concentrated likelihood is the ML estimate of σ^2 . The maximization is straightforward for the present case of the classical regression model, because $\mathbf{e}'\mathbf{e}$ is not a function of $\tilde{\sigma}^2$ and so can be taken as a constant. Still, taking the derivative with respect to $\tilde{\sigma}^2$, rather than with respect to $\tilde{\sigma}$, can be tricky. This can be avoided by denoting $\tilde{\sigma}^2$ by $\tilde{\gamma}$. Taking the derivative of (1.5.6) with respect to $\tilde{\gamma}$ ($\equiv \tilde{\sigma}^2$) and setting it to zero, we obtain the following result.

Proposition 1.5 (ML Estimator of $(\boldsymbol{\beta}, \sigma^2)$): *Suppose Assumptions 1.1–1.5 hold. Then the ML estimator of $\boldsymbol{\beta}$ is the OLS estimator \mathbf{b} and*

$$\text{ML estimator of } \sigma^2 = \frac{1}{n} \mathbf{e}'\mathbf{e} = \frac{SSR}{n} = \frac{n-K}{n} s^2. \quad (1.5.7)$$

We know from Proposition 1.2 that s^2 is unbiased. Since s^2 is multiplied by a factor $(n-K)/n$ which is different from 1, the ML estimator of σ^2 is biased, although the bias becomes arbitrarily small as the sample size n increases for any given fixed K .

For later use, we calculate the maximized value of the likelihood function. Substituting (1.5.7) into (1.5.6), we obtain

$$\text{maximized log likelihood} = -\frac{n}{2} \log\left(\frac{2\pi}{n}\right) - \frac{n}{2} - \frac{n}{2} \log(SSR),$$

so that the maximized likelihood is

$$\max_{\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2} L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) = \left(\frac{2\pi}{n}\right)^{-n/2} \cdot \exp\left(-\frac{n}{2}\right) \cdot (SSR)^{-n/2}. \quad (1.5.8)$$

Cramer-Rao Bound for the Classical Regression Model

Just to refresh your memory of basic statistics, we temporarily step outside the classical regression model and present without proof the Cramer-Rao inequality for the variance-covariance matrix of any unbiased estimator. For this purpose, define the **score vector** at a hypothetical parameter value $\tilde{\boldsymbol{\theta}}$ to be the gradient (vector of partial derivatives) of log likelihood:

$$\text{score: } \mathbf{s}(\tilde{\boldsymbol{\theta}}) \equiv \frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}. \quad (1.5.9)$$

Cramer-Rao Inequality: Let \mathbf{z} be a vector of random variables (not necessarily independent) the joint density of which is given by $f(\mathbf{z}; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is an m -dimensional vector of parameters in some parameter space Θ . Let $L(\tilde{\boldsymbol{\theta}}) \equiv f(\mathbf{z}; \tilde{\boldsymbol{\theta}})$ be the likelihood function, and let $\hat{\boldsymbol{\theta}}(\mathbf{z})$ be an unbiased estimator of $\boldsymbol{\theta}$ with a finite variance-covariance matrix. Then, under some regularity conditions on $f(\mathbf{z}; \boldsymbol{\theta})$ (not stated here),

$$\text{Var}[\hat{\boldsymbol{\theta}}(\mathbf{z})] \geq \mathbf{I}(\boldsymbol{\theta})^{-1} \quad (\equiv \text{Cramer-Rao Lower Bound}),$$

$(m \times m)$

where $\mathbf{I}(\boldsymbol{\theta})$ is the **information matrix** defined by

$$\mathbf{I}(\boldsymbol{\theta}) \equiv \text{E}[\mathbf{s}(\boldsymbol{\theta}) \mathbf{s}(\boldsymbol{\theta})']. \quad (1.5.10)$$

(Note well that the score is evaluated at the true parameter value $\boldsymbol{\theta}$.) Also under the regularity conditions, the information matrix equals the negative of the expected value of the Hessian (matrix of second partial derivatives) of the log likelihood:

$$\mathbf{I}(\boldsymbol{\theta}) = -\text{E} \left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\theta}}'} \right]. \quad (1.5.11)$$

This is called the **information matrix equality**.

See, e.g., Amemiya (1985, Theorem 1.3.1) for a proof and a statement of the regularity conditions. Those conditions guarantee that the operations of differentiation and taking expectations can be interchanged. Thus, for example,

$$\text{E}[\partial L(\boldsymbol{\theta}) / \partial \tilde{\boldsymbol{\theta}}] = \partial \text{E}[L(\boldsymbol{\theta})] / \partial \tilde{\boldsymbol{\theta}}.$$

Now, for the classical regression model (of Assumptions 1.1–1.5), the likelihood function $L(\tilde{\boldsymbol{\theta}})$ in the Cramer-Rao inequality is the conditional density (1.5.4), so the variance in the inequality is the variance conditional on \mathbf{X} . It can be shown that those regularity conditions are satisfied for the normal density (1.5.4) (see, e.g., Amemiya, 1985, Sections 1.3.2 and 1.3.3). In the rest of this subsection, we

calculate the information matrix for (1.5.4). The parameter vector $\boldsymbol{\theta}$ is $(\boldsymbol{\beta}', \sigma^2)'$. So $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\gamma})'$ and the matrix of second derivatives we seek to calculate is

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\frac{\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\theta}}'}{((K+1) \times (K+1))}} = \begin{bmatrix} \frac{\partial^2 \log L(\boldsymbol{\theta})}{\frac{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}'}{(K \times K)}} & \frac{\partial^2 \log L(\boldsymbol{\theta})}{\frac{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\gamma}}{(K \times 1)}} \\ \frac{\partial^2 \log L(\boldsymbol{\theta})}{\frac{\partial \tilde{\gamma} \partial \tilde{\boldsymbol{\beta}}'}{(1 \times K)}} & \frac{\partial^2 \log L(\boldsymbol{\theta})}{\frac{\partial^2 \tilde{\gamma}}{(1 \times 1)}} \end{bmatrix}. \quad (1.5.12)$$

The first and second derivatives of the log likelihood (1.5.5) with respect to $\tilde{\boldsymbol{\theta}}$, evaluated at the true parameter vector $\boldsymbol{\theta}$, are:

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\beta}}} = \frac{1}{\gamma} \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (1.5.13a)$$

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \tilde{\gamma}} = -\frac{n}{2\gamma^2} + \frac{1}{2\gamma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (1.5.13b)$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}'} = -\frac{1}{\gamma} \mathbf{X}'\mathbf{X}, \quad (1.5.14a)$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial^2 \tilde{\gamma}} = \frac{n}{2\gamma^2} - \frac{1}{\gamma^3} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (1.5.14b)$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\gamma}} = -\frac{1}{\gamma^2} \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (1.5.14c)$$

Since the derivatives are evaluated at the true parameter value, $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\varepsilon}$ in these expressions. Substituting (1.5.14) into (1.5.12) and using $E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$ (Assumption 1.2), $E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \mid \mathbf{X}) = n\sigma^2$ (implication of Assumption 1.4), and recalling $\gamma = \sigma^2$, we can easily derive

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma^2} \mathbf{X}'\mathbf{X} & \mathbf{0}' \\ \mathbf{0} & \frac{n}{2\sigma^4} \end{bmatrix}. \quad (1.5.15)$$

Here, the expectation is conditional on \mathbf{X} because the likelihood function (1.5.4) is a conditional density conditional on \mathbf{X} . This block diagonal matrix can be inverted to obtain the Cramer-Rao bound:

$$\text{Cramer-Rao bound} \equiv \mathbf{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2\sigma^4}{n} \end{bmatrix}. \quad (1.5.16)$$

Therefore, the unbiased estimator \mathbf{b} , whose variance is $\sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$ by Proposition 1.1, attains the Cramer-Rao bound. We have thus proved

Proposition 1.6 (b is the Best Unbiased Estimator (BUE)): *Under Assumptions 1.1–1.5, the OLS estimator \mathbf{b} of $\boldsymbol{\beta}$ is BUE (best unbiased estimator) in*

that any other unbiased (but not necessarily linear) estimator has smaller conditional variance in the matrix sense.

This result should be distinguished from the Gauss-Markov theorem that \mathbf{b} is minimum variance among those estimators that are unbiased *and* linear in \mathbf{y} . Proposition 1.6 says that \mathbf{b} is minimum variance in a larger class of estimators that includes nonlinear unbiased estimators. This stronger statement is obtained under the normality assumption (Assumption 1.5) which is not assumed in the Gauss-Markov theorem. Put differently, the Gauss-Markov theorem does not exclude the possibility of some nonlinear estimator beating OLS, but this possibility is ruled out by the normality assumption.

As was already seen, the ML estimator of σ^2 is biased, so the Cramer-Rao bound doesn't apply. But the OLS estimator s^2 of σ^2 is unbiased. Does it achieve the bound? We have shown in a review question to the previous section that

$$\text{Var}(s^2 | \mathbf{X}) = \frac{2\sigma^4}{n - K}$$

under the same set of assumptions as in Proposition 1.6. Therefore, s^2 does not attain the Cramer-Rao bound $2\sigma^4/n$. However, it can be shown that an unbiased estimator of σ^2 with variance lower than $2\sigma^4/(n - K)$ does not exist (see, e.g., Rao, 1973, p. 319).

The F -Test as a Likelihood Ratio Test

The **likelihood ratio test** of the null hypothesis compares L_U , the maximized likelihood without the imposition of the restriction specified in the null hypothesis, with L_R , the likelihood maximized subject to the restriction. If the likelihood ratio $\lambda \equiv L_U/L_R$ is too large, it should be a sign that the null is false. The F -test of the null hypothesis $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ considered in the previous section is a likelihood ratio test because the F -ratio is a monotone transformation of the likelihood ratio λ . For the present model, L_U is given by (1.5.8) where the SSR , the sum of squared residuals minimized without the constraint H_0 , is the SSR_U in (1.4.11). The restricted likelihood L_R is given by replacing this SSR by the restricted sum of squared residuals, SSR_R . So

$$L_R = \max_{\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2 \text{ s.t. } H_0} L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) = \left(\frac{2\pi}{n}\right)^{-n/2} \cdot \exp\left(-\frac{n}{2}\right) \cdot (SSR_R)^{-n/2}, \quad (1.5.17)$$

and the likelihood ratio is

$$\lambda \equiv \frac{L_U}{L_R} = \left(\frac{SSR_U}{SSR_R}\right)^{-n/2}. \quad (1.5.18)$$

Comparing this with the formula (1.4.11) for the F -ratio, we see that the F -ratio is a monotone transformation of the likelihood ratio λ :

$$F = \frac{n - K}{\# \mathbf{r}} (\lambda^{2/n} - 1), \quad (1.5.19)$$

so that the two tests are the same.

Quasi Maximum Likelihood

All these results assume the normality of the error term. Without normality, there is no guarantee that the ML estimator of β is OLS (Proposition 1.5) or that the OLS estimator \mathbf{b} achieves the Cramer-Rao bound (Proposition 1.6). However, Proposition 1.5 does imply that \mathbf{b} is a **quasi-** (or **pseudo-**) **maximum likelihood estimator**, an estimator that maximizes a mis-specified likelihood function. The mis-specified likelihood function we have considered is the normal likelihood. The results of Section 1.3 can then be interpreted as providing the finite sample properties of the quasi-ML estimator when the error is incorrectly specified to be normal.

Questions for Review

1. (Use of Regularity conditions) Assuming that taking expectations (i.e., taking integrals) and differentiation can be interchanged, prove that the expected value of the score vector given in (1.5.9), if evaluated at the true parameter value θ , is zero. **Hint:** What needs to be shown is that

$$\int \frac{\partial \log f(\mathbf{z}; \theta)}{\partial \tilde{\theta}} f(\mathbf{z}; \theta) d\mathbf{z} = \mathbf{0}.$$

Since $f(\mathbf{z}; \tilde{\theta})$ is a density, $\int f(\mathbf{z}; \tilde{\theta}) d\mathbf{z} = 1$ for any $\tilde{\theta}$. Differentiate both sides with respect to $\tilde{\theta}$ and use the regularity conditions, which allows us to change the order of integration and differentiation, to obtain: $\int [\partial f(\mathbf{z}; \theta) / \partial \tilde{\theta}] d\mathbf{z} = \mathbf{0}$. Also, from basic calculus,

$$\frac{\partial \log f(\mathbf{z}; \theta)}{\partial \tilde{\theta}} = \frac{1}{f(\mathbf{z}; \theta)} \frac{\partial f(\mathbf{z}; \theta)}{\partial \tilde{\theta}}.$$

2. (Maximizing joint log likelihood) Consider maximizing (the log of) the joint likelihood (1.5.2) for the classical regression model, where $\tilde{\theta} = (\tilde{\beta}', \tilde{\sigma}^2)'$ and $\log f(\mathbf{y} | \mathbf{X}; \tilde{\theta})$ is given by (1.5.5). You would parameterize the marginal likelihood $f(\mathbf{X}; \tilde{\psi})$ and take the log of (1.5.2) to obtain the objective function to

be maximized over $\zeta \equiv (\boldsymbol{\theta}', \boldsymbol{\psi}')'$. What is the ML estimator of $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \sigma^2)'$? [Answer: It should be the same as that in Proposition 1.5.] Derive the Cramer-Rao bound for $\boldsymbol{\beta}$. **Hint:** By the information matrix equality,

$$\mathbf{I}(\zeta) = -\mathbf{E} \left[\frac{\partial^2 \log L(\zeta)}{\partial \tilde{\zeta} \partial \tilde{\zeta}'} \right].$$

Also, $\partial^2 \log L(\zeta) / (\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\psi}}') = \mathbf{0}$.

3. (Concentrated log likelihood with respect to $\tilde{\sigma}^2$) Writing $\tilde{\sigma}^2$ as $\tilde{\gamma}$, the log likelihood function for the classical regression model is

$$\log L(\tilde{\boldsymbol{\beta}}, \tilde{\gamma}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\tilde{\gamma}) - \frac{1}{2\tilde{\gamma}} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}).$$

In the two-step maximization procedure described in the text, we first maximized this function with respect to $\tilde{\boldsymbol{\beta}}$. Instead, first maximize with respect to $\tilde{\gamma}$ given $\tilde{\boldsymbol{\beta}}$. Show that the concentrated log likelihood (concentrated with respect to $\tilde{\gamma} \equiv \tilde{\sigma}^2$) is

$$-\frac{n}{2} [1 + \log(2\pi)] - \frac{n}{2} \log \left(\frac{(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})}{n} \right).$$

4. (Information matrix equality for classical regression model) Verify (1.5.11) for the linear regression model.
5. (Likelihood equations for classical regression model) We used the two-step procedure to derive the ML estimate for the classical regression model. An alternative way to find the ML estimator is to solve for the first-order conditions which set (1.5.13) equal to zero (the first-order conditions for the log likelihood is called the **likelihood equations**). Verify that the ML estimator given in Proposition 1.5 solves the likelihood equations.

1.6 GLS (Generalized Least Squares)

Assumption 1.4 states that the $n \times n$ matrix of conditional second moments $\mathbf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}) (= \text{Var}(\boldsymbol{\varepsilon} | \mathbf{X}))$ is spherical, that is, proportional to the identity matrix. Without the assumption, each element of the $n \times n$ matrix is in general a nonlinear function of \mathbf{X} . If the error is not (conditionally) homoskedastic, the values of the diagonal elements of $\mathbf{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X})$ are not the same, and if there is correlation in

the error term between observations (the case of serial correlation for time-series models), the values of the off-diagonal elements are not zero. For any given positive scalar σ^2 , define $\mathbf{V}(\mathbf{X}) \equiv E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X})/\sigma^2$ and assume $\mathbf{V}(\mathbf{X})$ is nonsingular. That is,

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X}) = \sigma^2 \underset{(n \times n)}{\mathbf{V}(\mathbf{X})}, \quad \mathbf{V}(\mathbf{X}) \text{ nonsingular and known.} \quad (1.6.1)$$

The reason we decompose $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X})$ into the component σ^2 that is common to all elements of the matrix $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X})$ and the remaining component $\mathbf{V}(\mathbf{X})$ is that we don't need to know the value of σ^2 for efficient estimation. The model that results when Assumption 1.4 is replaced by (1.6.1), which merely assumes that the conditional second moment $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' | \mathbf{X})$ is nonsingular, is called the **generalized regression model**.

Consequence of Relaxing Assumption 1.4

Of the results derived in the previous sections, those that assume Assumption 1.4 are no longer valid for the generalized regression model. More specifically,

- The Gauss-Markov theorem no longer holds for the OLS estimator $\mathbf{b} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. The BLUE is some other estimator.
- The t -ratio is not distributed as the t distribution. Thus the t -test is no longer valid. The same comments apply to the F -test.
- However, the OLS estimator *is* still unbiased, because the unbiasedness result (Proposition 1.1(a)) does not require Assumption 1.4.

Efficient Estimation with Known \mathbf{V}

If the value of the matrix function $\mathbf{V}(\mathbf{X})$ is known, does there exist a BLUE for the generalized regression model? The answer is yes, and the estimator is called the **GLS (generalized least squares) estimator**, which we now derive. The basic idea of the derivation is to transform the generalized regression model, which consists of Assumptions 1.1–1.3 and (1.6.1), into a model that satisfies all the assumptions, including Assumption 1.4, of the classical regression model.

For economy of notation, we use \mathbf{V} for the value $\mathbf{V}(\mathbf{X})$. Since \mathbf{V} is by construction symmetric and positive definite, there exists a nonsingular $n \times n$ matrix \mathbf{C} such that

$$\mathbf{V}^{-1} = \mathbf{C}'\mathbf{C}. \quad (1.6.2)$$

This decomposition is not unique, with more than one choice for \mathbf{C} , but, as is clear from the discussion below, the choice of \mathbf{C} doesn't matter. Now consider creating

a new regression model by transforming $(\mathbf{y}, \mathbf{X}, \boldsymbol{\varepsilon})$ by \mathbf{C} as

$$\tilde{\mathbf{y}} \equiv \mathbf{C}\mathbf{y}, \quad \tilde{\mathbf{X}} \equiv \mathbf{C}\mathbf{X}, \quad \tilde{\boldsymbol{\varepsilon}} \equiv \mathbf{C}\boldsymbol{\varepsilon}. \quad (1.6.3)$$

Then Assumption 1.1 for $(\mathbf{y}, \mathbf{X}, \boldsymbol{\varepsilon})$ implies that $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\boldsymbol{\varepsilon}})$ too satisfies linearity:

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}}. \quad (1.6.4)$$

The transformed model satisfies the other assumptions of the classical linear regression model. Strict exogeneity is satisfied because

$$\begin{aligned} E(\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\mathbf{X}}) &= E(\tilde{\boldsymbol{\varepsilon}} \mid \mathbf{X}) \quad (\text{since } \mathbf{C} \text{ is nonsingular, } \mathbf{X} \text{ and } \tilde{\mathbf{X}} \text{ contain the same information}) \\ &= E(\mathbf{C}\boldsymbol{\varepsilon} \mid \mathbf{X}) \\ &= \mathbf{C} E(\boldsymbol{\varepsilon} \mid \mathbf{X}) \quad (\text{by the linearity of conditional expectations}) \\ &= \mathbf{0} \quad (\text{since } E(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0} \text{ by Assumption 1.2}). \end{aligned}$$

Because \mathbf{V} is positive definite, the no-multicollinearity assumption is also satisfied (see a review question below for a proof). Assumption 1.4 is satisfied for the transformed model because

$$\begin{aligned} E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}' \mid \tilde{\mathbf{X}}) &= E(\tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}' \mid \mathbf{X}) \quad (\text{since } \tilde{\mathbf{X}} \text{ and } \mathbf{X} \text{ contain the same information}) \\ &= \mathbf{C} E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X})\mathbf{C}' \quad (\text{since } \tilde{\boldsymbol{\varepsilon}}\tilde{\boldsymbol{\varepsilon}}' = \mathbf{C}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{C}') \\ &= \mathbf{C} \cdot \sigma^2 \cdot \mathbf{V}\mathbf{C}' \quad (\text{by (1.6.1)}) \\ &= \sigma^2\mathbf{C}\mathbf{V}\mathbf{C}' \\ &= \sigma^2\mathbf{I}_n \quad (\text{since } (\mathbf{C}')^{-1}\mathbf{V}^{-1}\mathbf{C}^{-1} = \mathbf{I}_n \text{ or } \mathbf{C}\mathbf{V}\mathbf{C}' = \mathbf{I}_n \text{ by (1.6.2)}). \end{aligned}$$

So indeed the variance of the transformed error vector $\tilde{\boldsymbol{\varepsilon}}$ is spherical. Finally, $\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\mathbf{X}}$ is normal because the distribution of $\tilde{\boldsymbol{\varepsilon}} \mid \tilde{\mathbf{X}}$ is the same as $\tilde{\boldsymbol{\varepsilon}} \mid \mathbf{X}$ and $\tilde{\boldsymbol{\varepsilon}}$ is a linear transformation of $\boldsymbol{\varepsilon}$. This completes the verification of Assumptions 1.1–1.5 for the transformed model.

The Gauss-Markov theorem for the transformed model implies that the BLUE of $\boldsymbol{\beta}$ for the generalized regression model is the OLS estimator applied to (1.6.4):

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{GLS}} &= (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{y}} \\ &= [(\mathbf{C}\mathbf{X})'(\mathbf{C}\mathbf{X})]^{-1}(\mathbf{C}\mathbf{X})'\mathbf{C}\mathbf{y} \\ &= (\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{C}'\mathbf{C}\mathbf{y}) \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad (\text{by (1.6.2)}). \end{aligned} \quad (1.6.5)$$

This is the GLS estimator. Its conditional variance is

$$\begin{aligned} &\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}} \mid \mathbf{X}) \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\text{Var}(\mathbf{y} \mid \mathbf{X})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}(\sigma^2\mathbf{V})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \quad (\text{since } \text{Var}(\mathbf{y} \mid \mathbf{X}) = \text{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})) \\ &= \sigma^2 \cdot (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \end{aligned} \quad (1.6.6)$$

Since replacing \mathbf{V} by $\sigma^2 \cdot \mathbf{V}$ ($= \text{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})$) in (1.6.5) doesn't change the numerical value, the GLS estimator can also be written as

$$\widehat{\boldsymbol{\beta}}_{\text{GLS}} = [\mathbf{X}' \text{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})^{-1} \mathbf{X}]^{-1} \mathbf{X}' \text{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})^{-1} \mathbf{y}.$$

As noted above, the OLS estimator $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ too is unbiased without Assumption 1.4, but nevertheless the GLS estimator should be preferred (provided \mathbf{V} is known) because the latter is more efficient in that the variance is smaller in the matrix sense. The gain in efficiency is achieved by exploiting the heteroskedasticity and correlation between observations in the error term, which, operationally, is to insert the inverse of (a matrix proportional to) $\text{Var}(\boldsymbol{\varepsilon} \mid \mathbf{X})$ in the OLS formula, as in (1.6.5). The discussion so far can be summarized as

Proposition 1.7 (finite-sample properties of GLS):

- (a) (unbiasedness) Under Assumption 1.1–1.3, $E(\widehat{\boldsymbol{\beta}}_{\text{GLS}} \mid \mathbf{X}) = \boldsymbol{\beta}$.
- (b) (expression for the variance) Under Assumptions 1.1–1.3 and the assumption (1.6.1) that the conditional second moment is proportional to $\mathbf{V}(\mathbf{X})$,

$$\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{GLS}} \mid \mathbf{X}) = \sigma^2 \cdot (\mathbf{X}'\mathbf{V}(\mathbf{X})^{-1}\mathbf{X})^{-1}.$$

- (c) (efficiency of GLS) Under the same set of assumptions as in (b), the GLS estimator is efficient in that the conditional variance of any unbiased estimator that is linear in \mathbf{y} is greater than or equal to $\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{GLS}} \mid \mathbf{X})$ in the matrix sense.

A Special Case: Weighted Least Squares (WLS)

The idea of adjusting for the error variance matrix becomes more transparent when there is no correlation in the error term between observations so that the matrix \mathbf{V} is diagonal. Let $v_i(\mathbf{X})$ be the i -th diagonal element of $\mathbf{V}(\mathbf{X})$. So

$$E(\varepsilon_i^2 \mid \mathbf{X}) (= \text{Var}(\varepsilon_i \mid \mathbf{X})) = \sigma^2 \cdot v_i(\mathbf{X}).$$

It is easy to see that \mathbf{C} is also diagonal, with the square root of $1/v_i(\mathbf{X})$ in the i -th diagonal. Thus $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}})$ is given by

$$\tilde{y}_i = \frac{y_i}{\sqrt{v_i(\mathbf{X})}}, \quad \tilde{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\sqrt{v_i(\mathbf{X})}} \quad (i = 1, 2, \dots, n).$$

Therefore, efficient estimation under a known form of heteroskedasticity is to first weight each observation by the reciprocal of the square root of the variance $v_i(\mathbf{X})$ and then apply OLS. This is called the **weighted regression** (or the **weighted least squares (WLS)**).

An important further special case is the case of a random sample where $\{y_i, \mathbf{x}_i\}$ is i.i.d. across i . As was noted in Section 1.1, the error is unconditionally homoskedastic (i.e., $E(\varepsilon_i^2)$ does not depend on i), but still GLS can be used to increase efficiency because the error can be conditionally heteroskedastic. The conditional second moment $E(\varepsilon_i^2 | \mathbf{X})$ for the case of random samples depends only on \mathbf{x}_i and the functional form of $E(\varepsilon_i^2 | \mathbf{x}_i)$ is the same across i . Thus

$$v_i(\mathbf{X}) = v(\mathbf{x}_i) \quad \text{for random samples.} \quad (1.6.7)$$

So the knowledge of $\mathbf{V}(\cdot)$ comes down to a single function of K variables, $v(\cdot)$.

Limiting Nature of GLS

All these sanguine conclusions about the finite-sample properties of GLS rest on the assumption that the regressors in the generalized regression model are strictly exogenous ($E(\tilde{\varepsilon} | \tilde{\mathbf{X}}) = \mathbf{0}$). This fact limits the usefulness of the GLS procedure. Suppose, as is often the case with time-series models, that the regressors are not strictly exogenous and the error is serially correlated. So neither OLS nor GLS has those good finite-sample properties such as unbiasedness. Nevertheless, as will be shown in the next chapter, the OLS estimator, which ignores serial correlation in the error, will have some good large sample properties (such as “consistency” and “asymptotic normality”), provided that the regressors are “predetermined” (which is weaker than strict exogeneity). The GLS estimator, in contrast, doesn’t have that redeeming feature. That is, if the error is not strictly exogenous but is merely predetermined, the GLS procedure to correct for serial correlation can make the estimator inconsistent. A procedure for explicitly taking serial correlation into account while maintaining consistency will be presented in Chapter 6.

If it is not appropriate for correcting for serial correlation, the GLS procedure can still be used to correct for heteroskedasticity when the error is not serially correlated with diagonal $\mathbf{V}(\mathbf{X})$, in the form of WLS. But that is provided that the matrix function $\mathbf{V}(\mathbf{X})$ is known. Very rarely do we have *a priori* information specifying the values of the diagonal elements of $\mathbf{V}(\mathbf{X})$, which is necessary to weight observations. In the case of a random sample where serial correlation is guaranteed not to arise, the knowledge of $\mathbf{V}(\mathbf{X})$ boils down to a single function of K variables, $v(\mathbf{x}_i)$, as we have just seen, but even for this case the knowledge of such a function is unavailable in most applications.

If we don’t know the function $\mathbf{V}(\mathbf{X})$, we can estimate its functional form from the sample. This approach is called the **Feasible Generalized Least Squares (FGLS)**. But if the function $\mathbf{V}(\mathbf{X})$ is estimated from the sample, its value \mathbf{V} becomes a random variable, which affects the distribution of the GLS estimator. Very little is known about the finite-sample properties of the FGLS estimator. We

will cover the large-sample properties of the FGLS estimator in the context of heteroskedasticity correction in the next chapter.

Before closing, one positive side of GLS should be noted: most linear estimation techniques — including the 2SLS, 3SLS, and the random effects estimators to be introduced later — can be expressed as a GLS estimator, with some liberal definition of data matrices. However, those estimators and OLS can also be interpreted as a GMM (generalized method of moments) estimator, and the GMM interpretation is more useful for developing large-sample results.

Questions for Review

1. (The no multi-collinearity assumption for the transformed model) Assumption 1.3 for the transformed model is that $\text{rank}(\mathbf{CX}) = K$. This is satisfied since \mathbf{C} is nonsingular and \mathbf{X} is of full column rank. Show this. **Hint:** Since \mathbf{X} is of full column rank, for any K -dimensional vector $\mathbf{c} \neq \mathbf{0}$, $\mathbf{Xc} \neq \mathbf{0}$.
2. (Generalized *SSR*) Show that $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ minimizes $(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})' \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$.
3. Derive the expression for $\text{Var}(\mathbf{b} \mid \mathbf{X})$ for the generalized regression model. What is the relation of it to $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}} \mid \mathbf{X})$? Verify that Proposition 1.7(c) implies

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \geq (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.$$

4. (sampling error of GLS) Show: $\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{\varepsilon}$.

1.7 Application: Returns to Scale in Electricity Supply

Nerlove's 1963 paper is a classic study of returns to scale in a regulated industry. It also is an excellent learning material for illustrating the techniques of this chapter and presenting a few more not yet covered.

The Electricity Supply Industry

At the time of Nerlove's writing, the U.S. electric power supply industry had the following features:

- (1) Privately owned local monopolies supply power on demand.
- (2) Rates (electricity prices) are set by the utility commission.
- (3) Factor prices (e.g. the wage rate) are given to the firm, either because of perfect competition in the market for factor inputs or through long-term contracts with labor unions.

These institutional features will be relevant when we examine whether the OLS is an appropriate estimation procedure.²³

The Data

Nerlove assembled a cross-section data set on 145 firms in 44 states in the year 1955 for which data on all the relevant variables are available. The variables in the data are: total costs, factor prices (the wage rate, the price of fuel, and the rental price of capital), and output. Although firms own capital (such as power plants, equipment, and structures), the standard investment theory of Jorgenson (1963) tells us that (as long as there is no costs in changing the capital stock) the firm should behave as if it rents capital on a period-to-period basis from itself at a rental price called the “user cost of capital”, which is defined as $(r + \delta) \cdot p_I$ where r here is the real interest rate (below we will use r for the degree of returns to scale), δ is the depreciation rate, and p_I is the price of capital goods. For this reason capital input can be treated as if it is a variable factor of production, just like labor and fuel inputs.

Appendix B of Nerlove (1963) contains a careful and honest discussion of how the data were constructed. Data on output, fuel and labor costs (which, along with capital costs, make up total costs) were obtained from Federal Power Commission (1956). For the wage rate Nerlove used state-wide average wages for utility workers. Ideally, one would calculate capital costs as the reproduction cost of capital times the user cost of capital. Due to data limitation, Nerlove instead used interest and depreciation charges available from the firm’s books.

Why Do We Need Econometrics?

Why do we need a fancy econometric technique like OLS to determine returns to scale? Why can’t we be simple-minded and plot the average cost (which can be easily calculated from the data as the ratio of total costs to output) against output and see whether the AC (average cost) curve is downward-sloping? The reason is

²³Thanks to the deregulation of the industry since the time of Nerlove’s writing, multiple firms are now allowed to compete in the same local market and the strict price control has been lifted in many states. So the first two features no longer characterize the industry.

that each firm can have a different AC curve. If firms face different factor prices, then the average cost is less for firms facing lower factor prices. That cross-section units at a given moment in time face the same prices is usually a good assumption to make, but not for the U.S. electricity industry with substantial regional differences in factor prices. The effect of factor prices on the AC curve has to be isolated somehow. The approach taken by Nerlove, which became a standard econometric practice, is to estimate a parameterized cost function.

Another factor that shifts the individual AC curve is the level of production efficiency. If more efficient firms produce more output, then it is possible that the individual AC curve is upward sloping but the line connecting the observed combination of the average cost and output is downward-sloping. To illustrate, consider a competitive industry described in Figure 1.6, where the AC and MC (marginal cost) curves are drawn for two firms competing in the same market. To focus on the connection between production efficiency and output, assume that all firms face the same factor prices so that the only reason the AC and MC curves differ between firms is the difference in production efficiency. The AC and MC curves are upward-sloping to reflect decreasing returns to scale. The AC and MC curves for firm A lie above those for firm B because firm A is less efficient than B. Because the industry is competitive, both firms face the same price p . Since output is determined at the intersection of the MC curve and the market price, the combinations of output and the average cost for two firms are points A and B in the figure. The curve obtained from connecting these two points can be downward-sloping, giving a false impression of *increasing* returns to scale.

The Cobb-Douglas Technology

To derive a parameterized cost function, we start with the Cobb-Douglas production function

$$Q_i = A_i x_{i1}^{\alpha_1} x_{i2}^{\alpha_2} x_{i3}^{\alpha_3}, \quad (1.7.1)$$

where Q_i is firm i 's output, x_{i1} is labor input for firm i , x_{i2} is capital input, and x_{i3} is fuel. A_i captures unobservable differences in production efficiency (this term is often called **firm heterogeneity**). The sum $\alpha_1 + \alpha_2 + \alpha_3 \equiv r$ is the degree of returns to scale. Thus, it is assumed *a priori* that the degree of returns to scale is constant (this should not be confused with constant returns to scale, which is that $r = 1$). Since the electric utilities in the sample are privately owned, it is reasonable to suppose that they are engaged in cost minimization (see, however, the discussion at the end of this section). We know from microeconomics that the cost function

associated with the Cobb-Douglas production function is Cobb-Douglas:

$$TC_i = r \cdot (A_i \alpha_1^{\alpha_1} \alpha_2^{\alpha_2} \alpha_3^{\alpha_3})^{-1/r} Q_i^{1/r} p_{i1}^{\alpha_1/r} p_{i2}^{\alpha_2/r} p_{i3}^{\alpha_3/r}, \quad (1.7.2)$$

where TC_i is total costs for firm i . Taking logs, we obtain the following log-linear relationship:

$$\log(TC_i) = \mu_i + \frac{1}{r} \log(Q_i) + \frac{\alpha_1}{r} \log(p_{i1}) + \frac{\alpha_2}{r} \log(p_{i2}) + \frac{\alpha_3}{r} \log(p_{i3}), \quad (1.7.3)$$

where $\mu_i = \log[r \cdot (A_i \alpha_1^{\alpha_1} \alpha_2^{\alpha_2} \alpha_3^{\alpha_3})^{-1/r}]$. The equation is said to be **log-linear** because both the dependent variable and the regressors are logs. Coefficients in log-linear equations are **elasticities**. The $\log(p_{i1})$ coefficient, for example, is the elasticity of total costs with respect to the wage rate, i.e., the percentage change in total costs when the wage rate changes by 1%. The degree of returns to scale, which in (1.7.3) is the reciprocal of the output elasticity of total costs, is independent of the level of output.

Now let $\mu \equiv E(\mu_i)$ and define $\varepsilon_i \equiv \mu_i - \mu$ so that $E(\varepsilon_i) = 0$. This ε_i represents the inverse of the firm's production efficiency relative to the industry's average efficiency; firms with positive ε_i are high-cost firms. With this notation, (1.7.3) becomes

$$\log(TC_i) = \beta_1 + \beta_2 \log(Q_i) + \beta_3 \log(p_{i1}) + \beta_4 \log(p_{i2}) + \beta_5 \log(p_{i3}) + \varepsilon_i. \quad (1.7.4)$$

where

$$\beta_1 = \mu \quad \beta_2 = \frac{1}{r}, \quad \beta_3 = \frac{\alpha_1}{r}, \quad \beta_4 = \frac{\alpha_2}{r}, \quad \text{and} \quad \beta_5 = \frac{\alpha_3}{r}. \quad (1.7.5)$$

Thus, the cost function has been cast in the regression format of Assumption 1.1 with $K = 5$. We noted a moment ago that the simple-minded approach of plotting the average cost against output cannot account for the factor price effect. What we have shown is that under the Cobb-Douglas technology the factor price effect is controlled for by the inclusion in the cost function of the logs of factor prices. Because the equation is derived from an explicit description of the firm's technology, the error term as well as the regression coefficients have clear interpretations.

How Do We Know Things are Cobb-Douglas?

The Cobb-Douglas functional form is certainly a very convenient parameterization of technology. But how do we know that the true production function is Cobb-Douglas? The Cobb-Douglas form satisfies the properties, such as diminishing marginal productivities, that we normally require for the production function, but the Cobb-Douglas form is certainly not the only functional form with those desirable

properties. A number of more general functional forms have been proposed in the literature, but the Cobb-Douglas form despite its simplicity has proved to be a surprisingly good description of technology. Nerlove's paper is one of the relatively few studies in which the Cobb-Douglas (log-linear) form is found to be inadequate, but it only underscores the importance of the Cobb-Douglas functional form as the benchmark from which one can usefully contemplate generalizations.

Are the OLS Assumptions Satisfied?

To justify the use of least squares, we need to make sure that Assumptions 1.1–1.4 are satisfied for the equation (1.7.4). Evidently, Assumption 1.1 (linearity) is satisfied with

$$y_i = \log(TC_i), \mathbf{x}_i = (1, \log(Q_i), \log(p_{i1}), \log(p_{i2}), \log(p_{i3}))'$$

There is no reason to expect that the regressors in (1.7.6) are perfectly multicollinear. Indeed, in Nerlove's data set, $\text{rank}(\mathbf{X}) = 5$ and $n = 145$, so Assumption 1.3 (no multi-collinearity) is satisfied as well.

In verifying the strict exogeneity assumption (Assumption 1.2), the features of the electricity industry mentioned above are relevant. It is reasonable to assume, as in most cross-section data, that \mathbf{x}_i is independent of ε_j for $i \neq j$. So the question is whether \mathbf{x}_i is independent of ε_i . If it is, then $E(\varepsilon | \mathbf{X}) = \mathbf{0}$. According to the third feature of the industry, factor prices are given to the firm with no regard for the firm's efficiency, so it is eminently reasonable to assume that factor prices are independent of ε_i .

What about output? Since the firm's output is supplied on demand (the first feature of the industry), output depends on the price of electricity set by the utility commission (the second feature). If the regulatory scheme is such that the price is determined regardless of the firm's efficiency, then $\log(Q_i)$ and ε_i are independently distributed. On the other hand, if the price is set to cover the average cost, then the firm's efficiency affects output through the effect of the electricity price on demand and output in this case is **endogenous**, being correlated with the error term. We will very briefly come back to this point at the end, but until then we will ignore the possible endogeneity of output. This certainly wouldn't do if we were dealing with a competitive industry. Since high-cost firms tend to produce less, there would be a *negative* correlation between $\log(Q_i)$ and ε_i , making OLS an inappropriate estimation procedure.

Regarding Assumption 1.4, the assumption of no correlation in the error term between firms (observations) would be suspect if, for example, there were technology spillovers running from one firm to other closely located firms. For the industry under study, this is probably not the case.

There is no *a priori* reason to suppose that homoskedasticity is satisfied. Indeed, the plot of residuals to be shown shortly suggests a failure of this condition. The main part of Nerlove's paper is exploring ways to deal with this problem.

Restricted Least Squares

The equation (1.7.4) is **over-identified** in that its five coefficients, being functions of the four technology parameters (which are α_1 , α_2 , α_3 , and μ), are not free parameters. We can easily see that from (1.7.5): $\beta_3 + \beta_4 + \beta_5 = 1$ (recall: $r \equiv \alpha_1 + \alpha_2 + \alpha_3$). This is a reflection of the generic property of the cost function that it is linearly homogeneous in factor prices. Indeed, multiplying total costs TC_i and all factor prices (p_{i1}, p_{i2}, p_{i3}) by a common factor leaves the cost function (1.7.4) intact if and only if $\beta_3 + \beta_4 + \beta_5 = 1$.

Estimating the equation by least squares while imposing *a priori* restrictions on the coefficient vector is the restricted least squares. It can be done easily by deriving from the original regression a separate regression that embodies the restrictions. In the present example, to impose the homogeneity restriction $\beta_3 + \beta_4 + \beta_5 = 1$ on the cost function, we take any one of the factor prices, say p_{i3} , and subtract $\log(p_{i3})$ from both sides of (1.7.4) to obtain

$$\log\left(\frac{TC_i}{p_{i3}}\right) = \beta_1 + \beta_2 \log(Q_i) + \beta_3 \log\left(\frac{p_{i1}}{p_{i3}}\right) + \beta_4 \log\left(\frac{p_{i2}}{p_{i3}}\right) + \varepsilon_i. \quad (1.7.6)$$

There are now four coefficients in the regression, from which unique values of the four technology parameters can be determined. The restricted least squares estimate of $(\beta_1, \dots, \beta_4)$ is simply the OLS estimate of the coefficients in (1.7.6). The restricted least squares estimate of β_5 is the value implied by the estimate of $(\beta_1, \dots, \beta_4)$ and the restriction.

Testing the Homogeneity of the Cost Function

Before proceeding to the estimation of the restricted model (1.7.6), in order to test the homogeneity restriction $\beta_3 + \beta_4 + \beta_5 = 1$, we first estimate the unrestricted model (1.7.4). If one uses the data available in printed form in Nerlove's paper, the OLS estimate of the equation is:

$$\log(TC_i) = -3.5 + 0.72 \log(Q_i) + 0.44 \log(p_{i1}) - 0.22 \log(p_{i2}) + 0.43 \log(p_{i3})$$

(1.8) (0.017) (0.29) (0.34) (0.10)

$$R^2 = 0.926, \text{ mean of dep. variable} = 1.72, \text{ SER} = 0.392, \text{ SSR} = 21.552, n = 145. \quad (1.7.7)$$

Here, numbers in parentheses are the standard errors of the OLS coefficient estimates. Since $\beta_2 = 1/r$, the estimate of the degree of returns to scale implied

by the OLS coefficient estimates is about 1.4 ($= 1/0.72$). The OLS estimate of $\beta_4 = \alpha_2/r$ has the wrong sign. As noted by Nerlove, there are reasons to believe that p_{i2} , the rental price of capital, is poorly measured. This may explain why b_4 is so imprecisely determined (i.e., the standard error is large relative to the size of the coefficient estimate) that one cannot reject the hypothesis that $\beta_4 = 0$ with a t -ratio of -0.65 ($= -0.22/0.34$).²⁴

To test the homogeneity restriction $H_0: \beta_3 + \beta_4 + \beta_5 = 1$, we could write the hypothesis in the form $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ with $\mathbf{R} = (0, 0, 1, 1, 1)$ and $\mathbf{r} = 1$, and use the formula (1.4.9) to calculate the F -ratio. The maintained hypothesis is the unrestricted model (1.7.4) (that is, Assumptions 1.1–1.5 where the equation in Assumption 1.1 is (1.7.4)), so the \mathbf{b} and the estimated variance of \mathbf{b} in the F -ratio formula should come from the OLS estimation of (1.7.4). Alternatively, we can use the F -ratio formula (1.4.11). The unrestricted model producing SSR_U is (1.7.4) and the restricted model producing SSR_R is (1.7.6) which superimposes the null hypothesis on the unrestricted model. The OLS estimate of (1.7.6) is:

$$\log\left(\frac{TC_i}{p_{i3}}\right) = \underset{(0.88)}{-4.7} + \underset{(0.017)}{0.72} \log(Q_i) + \underset{(0.20)}{0.59} \log(p_{i1}/p_{i3}) - \underset{(0.19)}{0.007} \log(p_{i2}/p_{i3})$$

$$R^2 = 0.932, \text{ mean of dep. var.} = -1.48, \text{ SER} = 0.39, \text{ SSR} = 21.640, n = 145. \quad (1.7.8)$$

The F test of the homogeneity restriction proceeds as follows.

Step 1: Using (1.4.11), the F -ratio can be calculated as

$$\frac{(21.640 - 21.552)/1}{21.552/(145 - 5)} = 0.57.$$

Step 2: Find the critical value. The number of restrictions (equations) in the null hypothesis is 1 and K (the number of coefficients) in the unrestricted model (which is the maintained hypothesis) is 5. So the degrees of freedom are 1 and 140 ($= 145 - 5$). From the table of F distributions, the critical value is about 3.9.

Step 3: Thus, we can easily accept the homogeneity restriction, a very comforting conclusion for those who take microeconomics seriously (like us).

²⁴The consequence of measurement error is not just that the coefficient of the variable measured with error is poorly determined; it could also contaminate the coefficient estimates for all other regressors. The appropriate context to address this problem is the large sample theory for endogenous regressors in Chapter 3.

Detour: A Cautionary Note on R^2

The R^2 of 0.926 is surprisingly high for cross-section estimates, but some of the explanatory power of the regression comes from the scale effect that total costs increase with firm size. To gauge the contribution of the scale effect on the R^2 , subtract $\log(Q_i)$ from both sides of (1.7.4) to obtain an equivalent cost function:

$$\log\left(\frac{TC_i}{Q_i}\right) = \beta_1 + (\beta_2 - 1)\log(Q_i) + \beta_3 \log(p_{i1}) + \beta_4 \log(p_{i2}) + \beta_5 \log(p_{i3}) + \varepsilon_i. \quad (1.7.4')$$

Here, the dependent variable is the average cost rather than total costs. Application of the OLS to (1.7.4') using the same data yields:

$$\log\left(\frac{TC_i}{Q_i}\right) = \begin{matrix} -3.5 \\ (1.8) \end{matrix} - \begin{matrix} 0.28 \log(Q_i) \\ (0.017) \end{matrix} + \begin{matrix} 0.44 \log(p_{i1}) \\ (0.29) \end{matrix} - \begin{matrix} 0.22 \log(p_{i2}) \\ (0.34) \end{matrix} + \begin{matrix} 0.43 \log(p_{i3}) \\ (0.10) \end{matrix}$$

$$R^2 = 0.695, \text{ mean of dep. var.} = -4.83, \text{ SER} = 0.392, \text{ SSR} = 21.552, n = 145. \quad (1.7.9)$$

As you no doubt have anticipated, the output coefficient is now -0.28 ($= 0.72 - 1$) with the standard errors and the other coefficient estimates unchanged. The R^2 changes only because the dependent variable is different. It is nonsense to say that the higher R^2 makes (1.7.4) preferable to (1.7.4'), because the two equations represent the same model. The point is: when comparing equations on the basis of the fit, the equations must share the same dependent variable.

Testing Constant Returns to Scale

As an application of the t -test, consider testing whether returns to scale are constant ($r = 1$). We take the maintained hypothesis to be the restricted model (1.7.6). Because β_2 (the log output coefficient) equals 1 if and only if $r = 1$, the null hypothesis is that $H_0: \beta_2 = 1$. The t test of constant returns to scale proceeds as follows.

Step 1: Calculate the t -ratio for the hypothesis. From the estimation of the restricted model, we have $b_2 = 0.72$ with a standard error of 0.017, so

$$t\text{-ratio} = \frac{0.72 - 1}{0.017} = -16.$$

Because the maintained hypothesis here is the restricted model (1.7.6), K (the number of coefficients) = 4.

Step 2: Look for the critical value in the $t(141)$ distribution. If the size of the test is 5%, the critical value is 1.98.

Step 3: Since the absolute value of the t -ratio is far greater than the critical value, we reject the hypothesis of constant returns to scale.

Importance of Plotting Residuals

The regression has a problem which cannot be seen from the estimated coefficients and their standard errors. Figure 1.7 plots the residuals against $\log(Q_i)$. Notice two things from the plot. First, as output increases, the residuals first tend to be positive, then negative, and again positive. This strongly suggests that the degree of returns to scale (r) is not constant as assumed in the log-linear specification. Second, the residuals are more widely scattered for lower outputs, which is a sign of a failure of the homoskedasticity assumption that the error variance doesn't depend on the regressors. To deal with these problems, Nerlove divided the sample of 145 firms into five groups of 29, ordered by output, and estimated the model (1.7.6) separately for each group. This amounts to allowing all the coefficients (including $\beta_2 = 1/r$) and the error variance to differ across the five groups differing in size. Nerlove finds that returns to scale diminish steadily, from a high of well over 2 to a low of slightly below 1, over the output range of the data. In the empirical exercise of this chapter, the reader is asked to replicate this finding and do some further analysis using **dummy variables** and the weighted least squares.

Subsequent Developments

One strand of the subsequent literature is concerned about generalizing the Cobb-Douglas technology while maintaining the assumption of cost minimization. An obvious alternative to Cobb-Douglas is the CES (Constant Elasticity of Substitution) production function, but it has two problems. First, the cost function implied by the CES production function is highly nonlinear (which, though, could be overcome by the use of nonlinear least squares to be covered in Chapter 7). Second, the CES technology implies a constant degree of returns to scale. One of Nerlove's main findings is that the degree varies with output. Christensen and Greene (1976) is probably the first to estimate the technology parameters allowing for variable degrees of returns to scale. Using the **translog cost function** introduced by Christensen, Jorgenson, and Lau (1973), they find that the significant scale economies evident in the 1955 data were mostly exhausted by 1970, with most firms operating at much higher output levels where the AC curve is essentially flat. Their work will be examined in detail in Chapter 4.

Another issue is whether regulated firms minimize costs. The influential paper by Averch and Johnson (1962) argues that the practice by regulators to guarantee utilities a "fair rate of return" on their capital stock distorts the choice of input

levels. Since the fair rate of return is usually higher than the interest rate, utilities have an incentive to over-invest. That is, they minimize costs but the relevant rate of return in the definition of the user cost of capital is the fair rate of return. Consequently, unless the fair rate of return is used in the calculation of p_{i2} , the true technology parameters cannot be estimated from the cost function. The fair-rate-of-return regulation creates another econometric problem: to guarantee utilities a fair rate of return, the price of electricity must be kept relatively high in markets served by high-cost utilities. Thus output will be endogenous.

A more recent issue is whether the regulator has enough information to bring about cost minimization. If the utility has more information about costs, it has an incentive to misreport to the regulator the true value of the efficiency parameter. Schemes to be adopted by the regulator to take into account this incentive problem may not lead to cost minimization. Wolak's (1994) empirical results for California's water utility industry indicate that the observed level of costs and output is better modeled as the outcome of a regulator-utility interaction under asymmetric information. Wolak resolves the problem of the endogeneity of output by estimating the demand function along with the cost function. Doing so, however, requires an estimation technique more sophisticated than the OLS.

Questions for Review

1. (Review of duality theory) Consult your favorite micro textbook to remember how to derive the Cobb-Douglas cost function from the Cobb-Douglas production function.
2. (Change of units) In Nerlove's data, output is measured in kilowatt hours. If output were measured in megawatt hours, how would the estimated restricted regression change?
3. (Recovering technology parameters from regression coefficients) Show that the technology parameters $(\mu, \alpha_1, \alpha_2, \alpha_3)$ can be determined uniquely from the first four equations in (1.7.5) and the definition $r \equiv \alpha_1 + \alpha_2 + \alpha_3$. (Don't use the fifth equation $\beta_5 = \alpha_3/r$.)
4. (Recovering left-out coefficients from restricted OLS) Calculate the restricted OLS estimate of β_5 from (1.7.8). How do you calculate the standard error of b_5 from the printout of the restricted OLS? **Hint:** Write $b_5 = a + \mathbf{c}'\mathbf{b}$ for suitably chosen a and \mathbf{c} where \mathbf{b} here is $(b_1, \dots, b_4)'$. So $\text{Var}(b_5 | \mathbf{X}) = \mathbf{c}' \text{Var}(\mathbf{b} | \mathbf{X}) \mathbf{c}$. The printout from the restricted OLS should include $\text{Var}(\widehat{\mathbf{b}} | \mathbf{X})$.

5. If you take p_{i2} instead of p_{i3} and subtract $\log(p_{i2})$ from both sides of (1.7.7), how does the restricted regression look? Without actually estimating it on Nerlove's data, can you tell from the estimated restricted regression in the text what the restricted OLS estimate of $(\beta_1, \dots, \beta_5)$ will be? Their standard errors? The *SSR*? What about the R^2 ?
6. Why is the R^2 of 0.926 from the unrestricted model (1.7.7) *lower* than the R^2 of 0.932 from the restricted model (1.7.8)?
7. A more realistic assumption about the rental price of capital may be that there is an economy-wide capital market so p_{i2} is the same across firms. In this case,
 - (a) Can we estimate the technology parameters? **Hint:** The answer is yes, but why? When p_{i2} is constant, (1.7.4) will have the perfect multicollinearity problem. But recall that $(\beta_1, \dots, \beta_5)$ are not free parameters.
 - (b) Can we test homogeneity of the cost function in factor prices?
8. Taking logs of both sides of the production function (1.7.1), one can derive the log-linear relationship:

$$\log(Q_i) = \alpha_0 + \alpha_1 \log(x_{i1}) + \alpha_2 \log(x_{i2}) + \alpha_3 \log(x_{i3}) + \varepsilon_i,$$

where ε_i here is defined as $\log(A_i) - E[\log(A_i)]$ and $\alpha_0 = E[\log(A_i)]$. Suppose, in addition to total costs, output, and factor prices, you had data on factor inputs. Can we estimate α 's by applying OLS to this log-linear relationship? Why or why not? **Hint:** Do input levels depend on ε_i ? Suggest a different way to estimate α 's. **Hint:** Look at input shares.

Problem Set for Chapter 1

Analytical Exercises

1. (Proof that \mathbf{b} minimizes SSR) Let \mathbf{b} be the OLS estimator of $\boldsymbol{\beta}$. Prove that, for any hypothetical estimate, $\tilde{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$,

$$(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \geq (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}).$$

In your proof, use the “add-and-subtract” strategy: take $\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$, add $\mathbf{X}\mathbf{b}$ to it and then subtract the same from it. It produces the decomposition of $\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}$:

$$\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}} = (\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{X}\mathbf{b} - \mathbf{X}\tilde{\boldsymbol{\beta}}).$$

Hint: $(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = [(\mathbf{y} - \mathbf{X}\mathbf{b}) + \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}})]'[(\mathbf{y} - \mathbf{X}\mathbf{b}) + \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}})]$. Using the normal equations, show that this equals

$$(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \tilde{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\mathbf{b} - \tilde{\boldsymbol{\beta}}).$$

2. (The annihilator associated with the vector of ones) Let $\mathbf{1}$ be the n -dimensional column vector of ones and let $\mathbf{M}_1 \equiv \mathbf{I}_n - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$. That is, \mathbf{M}_1 is the annihilator associated with $\mathbf{1}$. Prove the following.

- (a) \mathbf{M}_1 is symmetric and idempotent.
 (b) $\mathbf{M}_1\mathbf{1} = \mathbf{0}$.
 (c) $\mathbf{M}_1\mathbf{y} = \mathbf{y} - \bar{y} \cdot \mathbf{1}$ where

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

$\mathbf{M}_1\mathbf{y}$ is the vector of **deviations from the mean**.

- (d) $\mathbf{M}_1\mathbf{X} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$ where $\bar{\mathbf{x}} = \mathbf{X}'\mathbf{1}/n$. The k -th element of the $K \times 1$ vector $\bar{\mathbf{x}}$ is $\frac{1}{n} \sum_{i=1}^n x_{ik}$.

3. (Deviation-from-the-mean regression) Consider a regression model with a constant. Let \mathbf{X} be partitioned as

$$\underset{(n \times K)}{\mathbf{X}} = \begin{bmatrix} \mathbf{1} & \vdots & \mathbf{X}_2 \\ n \times 1 & & n \times (K-1) \end{bmatrix}$$

So the first regressor is a constant. Partition $\boldsymbol{\beta}$ and \mathbf{b} accordingly:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \begin{array}{l} \leftarrow \text{scalar} \\ \leftarrow (K-1) \times 1 \end{array}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \mathbf{b}_2 \end{bmatrix}.$$

Also let $\tilde{\mathbf{X}}_2 \equiv \mathbf{M}_1 \mathbf{X}_2$ and $\tilde{\mathbf{y}} \equiv \mathbf{M}_1 \mathbf{y}$. They are the deviations from the mean for the non-constant regressors and the dependent variable. Prove the following.

(a) The K normal equations are:

$$\bar{y} - b_1 - \bar{\mathbf{x}}_2' \mathbf{b}_2 = 0$$

where $\bar{\mathbf{x}}_2 = \mathbf{X}_2' \mathbf{1} / n$,

$$\mathbf{X}_2' \mathbf{y} - n \cdot b_1 \cdot \bar{\mathbf{x}}_2 - \mathbf{X}_2' \mathbf{X}_2 \mathbf{b}_2 = \begin{matrix} \mathbf{0} \\ ((K-1) \times 1) \end{matrix}.$$

(b) $\mathbf{b}_2 = (\tilde{\mathbf{X}}_2' \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2' \tilde{\mathbf{y}}$. **Hint:** Substitute the first normal equation into the other $K-1$ equations to eliminate b_1 and solve for \mathbf{b}_2 . This is a generalization of the result you proved in Review Question 3 in Section 1.2.

4. (Partitioned regression, generalization of Exercise 3) Let \mathbf{X} be partitioned as

$$\underset{(n \times K)}{\mathbf{X}} = \begin{bmatrix} \underset{(n \times K_1)}{\mathbf{X}_1} & \vdots & \underset{(n \times K_2)}{\mathbf{X}_2} \end{bmatrix}$$

Partition $\boldsymbol{\beta}$ accordingly:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \begin{array}{l} \leftarrow K_1 \times 1 \\ \leftarrow K_2 \times 1 \end{array}.$$

Thus the regression can be written as

$$\mathbf{y} = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}.$$

Let $\mathbf{P}_1 \equiv \mathbf{X}_1 (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1'$, $\mathbf{M}_1 \equiv \mathbf{I} - \mathbf{P}_1$, $\tilde{\mathbf{X}}_2 \equiv \mathbf{M}_1 \mathbf{X}_2$ and $\tilde{\mathbf{y}} \equiv \mathbf{M}_1 \mathbf{y}$. Thus, $\tilde{\mathbf{y}}$ is the residual vector from the regression of \mathbf{y} on \mathbf{X}_1 , and the k -th column of $\tilde{\mathbf{X}}_2$ is the residual vector from the regression of the corresponding k -th column of \mathbf{X}_2 on \mathbf{X}_1 . Prove the following.

(a) The normal equations are:

$$\mathbf{X}_1' \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_1' \mathbf{X}_2 \boldsymbol{\beta}_2 = \mathbf{X}_1' \mathbf{y}, \quad (*)$$

$$\mathbf{X}_2' \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2' \mathbf{X}_2 \boldsymbol{\beta}_2 = \mathbf{X}_2' \mathbf{y}. \quad (**)$$

- (b) $\mathbf{b}_2 = (\tilde{\mathbf{X}}_2' \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2' \tilde{\mathbf{y}}$. That is, \mathbf{b}_2 can be obtained by regressing the residuals $\tilde{\mathbf{y}}$ on the matrix of residuals $\tilde{\mathbf{X}}_2$. **Hint:** Derive $\mathbf{X}_1 \beta_1 = -\mathbf{P}_1 \mathbf{X}_2 \beta_2 + \mathbf{P}_1 \mathbf{y}$ from (*). Substitute this into (**) to obtain: $\mathbf{X}_2' \mathbf{M}_1 \mathbf{X}_2 \beta_2 = \mathbf{X}_2' \mathbf{M}_1 \mathbf{y}$. Then use the fact that \mathbf{M}_1 is symmetric and idempotent. Or, if you wish, you can apply the brute force of the partitioned inverse formula (A.10) of Appendix A to the coefficient matrix

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}'_1 \mathbf{X}_1 & \mathbf{X}'_1 \mathbf{X}_2 \\ \mathbf{X}'_2 \mathbf{X}_1 & \mathbf{X}'_2 \mathbf{X}_2 \end{bmatrix}.$$

Show that the second diagonal block of $(\mathbf{X}'\mathbf{X})^{-1}$ is $(\tilde{\mathbf{X}}_2' \tilde{\mathbf{X}}_2)^{-1}$.

- (c) The residuals from the regression of $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}_2$ numerically equals \mathbf{e} , the residuals from the regression of \mathbf{y} on \mathbf{X} ($\equiv (\mathbf{X}_1 : \mathbf{X}_2)$). **Hint:** If \mathbf{e} is the residual from the regression of \mathbf{y} on \mathbf{X} ,

$$\mathbf{y} = \mathbf{X}_1 \mathbf{b}_1 + \mathbf{X}_2 \mathbf{b}_2 + \mathbf{e}.$$

Pre-multiplying both sides by \mathbf{M}_1 and using $\mathbf{M}_1 \mathbf{X}_1 = \mathbf{0}$, we obtain

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_2 \mathbf{b}_2 + \mathbf{M}_1 \mathbf{e}.$$

Show that $\mathbf{M}_1 \mathbf{e} = \mathbf{e}$ and observe that \mathbf{b}_2 equals the OLS coefficient estimate in the regression of $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}_2$.

- (d) $\mathbf{b}_2 = (\tilde{\mathbf{X}}_2' \tilde{\mathbf{X}}_2)^{-1} \tilde{\mathbf{X}}_2' \mathbf{y}$. Note the difference from (b). Here, the vector of dependent variable is \mathbf{y} , not $\tilde{\mathbf{y}}$. Are the residuals from the regression of \mathbf{y} on $\tilde{\mathbf{X}}_2$ numerically the same as \mathbf{e} ? [Answer: No.] Is the *SSR* from the regression of \mathbf{y} on $\tilde{\mathbf{X}}_2$ the same as the *SSR* from the regression of $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}_2$? [Answer: No.]

The results in (b)–(d) are known as the **Frisch-Waugh theorem**.

- (e) Show:

$$\tilde{\mathbf{y}}' \tilde{\mathbf{y}} - \mathbf{e}' \mathbf{e} = \tilde{\mathbf{y}}' \mathbf{X}_2 (\mathbf{X}_2' \mathbf{M}_1 \mathbf{X}_2)^{-1} \mathbf{X}_2' \tilde{\mathbf{y}}.$$

Hint: Apply the general decomposition formula (1.2.15) to the regression in (c) to derive

$$\tilde{\mathbf{y}}' \tilde{\mathbf{y}} = \mathbf{b}_2' \tilde{\mathbf{X}}_2' \tilde{\mathbf{X}}_2 \mathbf{b}_2 + \mathbf{e}' \mathbf{e}.$$

Then use (b).

- (f) Consider the following four regressions:

- (1) regress $\tilde{\mathbf{y}}$ on \mathbf{X}_1 .

- (2) regress $\tilde{\mathbf{y}}$ on $\tilde{\mathbf{X}}_2$.
 (3) regress $\tilde{\mathbf{y}}$ on \mathbf{X}_1 and \mathbf{X}_2 .
 (4) regress $\tilde{\mathbf{y}}$ on \mathbf{X}_2 .

Let SSR_j be the sum of squared residuals from regression j . Show:

- (i) $SSR_1 = \tilde{\mathbf{y}}'\tilde{\mathbf{y}}$. **Hint:** $\tilde{\mathbf{y}}$ is constructed so that $\mathbf{X}'_1\tilde{\mathbf{y}} = \mathbf{0}$, so \mathbf{X}_1 should have no explanatory power.
 (ii) $SSR_2 = \mathbf{e}'\mathbf{e}$. **Hint:** Use (c).
 (iii) $SSR_3 = \mathbf{e}'\mathbf{e}$. **Hint:** Apply the Frisch-Waugh theorem on regression (3). $\mathbf{M}_1\tilde{\mathbf{y}} = \tilde{\mathbf{y}}$.
 (iv) Verify by numerical example that SSR_4 is not necessarily equal to $\mathbf{e}'\mathbf{e}$.
5. (Restricted regression and F) In the restricted least squares, the sum of squared residuals is minimized subject to the constraint implied by the null hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. Form the Lagrangian as

$$\mathcal{L} = \frac{1}{2}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) + \boldsymbol{\lambda}'(\mathbf{R}\tilde{\boldsymbol{\beta}} - \mathbf{r}),$$

where $\boldsymbol{\lambda}$ here is the $\#\mathbf{r}$ -dimensional vector of Lagrange multipliers (recall: \mathbf{R} is $\#\mathbf{r} \times K$, $\tilde{\boldsymbol{\beta}}$ is $K \times 1$, and \mathbf{r} is $\#\mathbf{r} \times 1$). Let $\hat{\boldsymbol{\beta}}$ be the restricted least squares estimator of $\boldsymbol{\beta}$. It is the solution to the constrained minimization problem.

- (a) Let \mathbf{b} be the unrestricted OLS estimator. Show:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}), \\ \boldsymbol{\lambda} &= [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}).\end{aligned}$$

Hint: The first-order conditions are: $\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{R}'\boldsymbol{\lambda}$ or $\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{R}'\boldsymbol{\lambda}$. Combine this with the constraint $\mathbf{R}\hat{\boldsymbol{\beta}} = \mathbf{r}$ to solve for $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\beta}}$.

- (b) Let $\hat{\boldsymbol{\varepsilon}} \equiv \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, the residuals from the restricted regression. Show:

$$\begin{aligned}SSR_R - SSR_U &= (\mathbf{b} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\mathbf{b} - \hat{\boldsymbol{\beta}}) \\ &= (\mathbf{R}\mathbf{b} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) \\ &= \boldsymbol{\lambda}'\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda} \\ &= \hat{\boldsymbol{\varepsilon}}'\mathbf{P}\hat{\boldsymbol{\varepsilon}}.\end{aligned}$$

where \mathbf{P} is the projection matrix. **Hint:** For the first equality, use the “add-and-subtract” strategy:

$$SSR_R = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = [(\mathbf{y} - \mathbf{X}\mathbf{b}) + \mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}})]'[(\mathbf{y} - \mathbf{X}\mathbf{b}) + \mathbf{X}(\mathbf{b} - \hat{\boldsymbol{\beta}})].$$

Use the normal equations $\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0}$. For the second and third equalities, use (a). To prove the fourth equality, the easiest way is to use the first-order condition mentioned in (a) that $\mathbf{R}'\boldsymbol{\lambda} = \mathbf{X}'\hat{\boldsymbol{\varepsilon}}$.

(c) Verify that you have proved in (b) that (1.4.9) = (1.4.11).

6. (proof of the decomposition (1.2.17)) Take the unrestricted model to be a regression where one of the regressors is a constant, and the restricted model to be a regression where the only regressor is a constant.

(a) Show that (b) in the previous exercise is the decomposition (1.2.17) for this case. **Hint:** What is $\hat{\boldsymbol{\beta}}$ for this case? Show that $SSR_R = \sum_i (y_i - \bar{y})^2$ and $(\mathbf{b} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\mathbf{b} - \hat{\boldsymbol{\beta}}) = \sum_i (\hat{y}_i - \bar{y})^2$.

(b) (R^2 as an F -ratio) For a regression where one of the regressors is a constant, prove that

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}.$$

7. (Hausman principle in finite samples) For the generalized regression model, prove the following. Here, it is understood that the expectations, variances, and covariances are all conditional on \mathbf{X} .

(a) $\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{GLS}}, \mathbf{b} - \hat{\boldsymbol{\beta}}_{\text{GLS}}) = \mathbf{0}$. **Hint:** Recall that, for any two random vectors \mathbf{x} and \mathbf{y} ,

$$\text{Cov}(\mathbf{x}, \mathbf{y}) \equiv \text{E}[(\mathbf{x} - \text{E}(\mathbf{x}))(\mathbf{y} - \text{E}(\mathbf{y}))'].$$

So

$$\text{Cov}(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{y}) = \mathbf{A} \text{Cov}(\mathbf{x}, \mathbf{y}) \mathbf{B}'.$$

Also, since $\boldsymbol{\beta}$ is non-random,

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{GLS}}, \mathbf{b} - \hat{\boldsymbol{\beta}}_{\text{GLS}}) = \text{Cov}(\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta}, \mathbf{b} - \hat{\boldsymbol{\beta}}_{\text{GLS}}).$$

(b) Let $\tilde{\boldsymbol{\beta}}$ be any unbiased estimator and define $\mathbf{q} \equiv \tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{GLS}}$. Assume $\tilde{\boldsymbol{\beta}}$ is such that $\mathbf{V}_{\mathbf{q}} \equiv \text{Var}(\mathbf{q})$ is nonsingular. Prove: $\text{Cov}(\hat{\boldsymbol{\beta}}_{\text{GLS}}, \mathbf{q}) = \mathbf{0}$. (If we set $\tilde{\boldsymbol{\beta}} = \mathbf{b}$, we are back to (a).) **Hint:** Define: $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}_{\text{GLS}} + \mathbf{H}\mathbf{q}$ for some \mathbf{H} . Show:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) + \mathbf{C}\mathbf{H}' + \mathbf{H}\mathbf{C}' + \mathbf{H}\mathbf{V}_{\mathbf{q}}\mathbf{H}',$$

where $\mathbf{C} \equiv \text{Cov}(\hat{\boldsymbol{\beta}}_{\text{GLS}}, \mathbf{q})$. Show that, if $\mathbf{C} \neq \mathbf{0}$ then $\text{Var}(\hat{\boldsymbol{\beta}})$ can be made smaller than $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}})$ by setting $\mathbf{H} = -\mathbf{C}\mathbf{V}_{\mathbf{q}}^{-1}$. Argue that this is in contradiction to Proposition 1.7(c).

- (c) (optional, only for those who are proficient in linear algebra) Prove: if the K columns of \mathbf{X} are characteristic vectors of \mathbf{V} , then $\mathbf{b} = \hat{\boldsymbol{\beta}}_{\text{GLS}}$, where \mathbf{V} is the $n \times n$ variance-covariance matrix of the n -dimensional error vector $\boldsymbol{\varepsilon}$. (So not all unbiased estimators satisfy the requirement in (b) that $\text{Var}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{\text{GLS}})$ be nonsingular.) **Hint:** For any $n \times n$ symmetric matrix \mathbf{V} , there exists an $n \times n$ matrix \mathbf{H} such that $\mathbf{H}'\mathbf{H} = \mathbf{I}_n$ (so \mathbf{H} is an orthogonal matrix) and $\mathbf{H}'\mathbf{V}\mathbf{H} = \boldsymbol{\Lambda}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix with the characteristic roots (which are real since \mathbf{V} is symmetric) of \mathbf{V} in the diagonal. The columns of \mathbf{H} are called the characteristic vectors of \mathbf{V} . Show that

$$\mathbf{H}^{-1} = \mathbf{H}', \quad \mathbf{H}'\mathbf{V}^{-1}\mathbf{H} = \boldsymbol{\Lambda}^{-1}, \quad \mathbf{H}'\mathbf{V}^{-1} = \boldsymbol{\Lambda}^{-1}\mathbf{H}'.$$

Without loss of generality, \mathbf{X} can be taken to be the first K columns of \mathbf{H} . So $\mathbf{X} = \mathbf{H}\mathbf{F}$, where

$$\underset{(n \times K)}{\mathbf{F}} = \begin{bmatrix} \mathbf{I}_K \\ \mathbf{0} \end{bmatrix}.$$

Empirical Exercise

(Read Marc Nerlove, “Returns to Scale in Electricity Supply” (except paragraphs of equations (6)–(9), the part of section 2 from p. 184 on, and Appendix A and C) before doing this exercise.) For 145 electric utility companies in 1955, the file NERLOVE.ASC has data on the following:

Column 1: total costs (call it TC) in millions of dollars,

Column 2: output (Q) in billions of kilowatt hours,

Column 3: price of labor (PL),

Column 4: price of fuels (PF),

Column 5: price of capital (PK).

They are from the data appendix of his article. There are 145 observations, and the observations are ordered in size, observation 1 being the smallest company and observation 145 the largest. Using the data transformation facilities of your computer software, generate for each of the 145 firms the variables required for estimation. To estimate (1.7.4), for example, you need to generate $\log(TC)$, a constant, $\log(Q)$, $\log(PL)$, $\log(PK)$, and $\log(PF)$, for each of the 145 firms.

- (a) (data question) Does Nerlove’s construction of the price of capital conform to the definition of the user cost of capital? **Hint:** Read Nerlove’s Appendix B.4.
- (b) Estimate the unrestricted model (1.7.4) by OLS. Can you replicate the estimates in the text?
- (c) (Restricted least squares) Estimate the restricted model (1.7.6) by OLS. To do this, you need to generate a new set of variables for each of the 145 firms. For example, the dependent variable is $\log(TC/PF)$, not $\log(TC)$. Can you replicate the estimates in the text? Can you replicate Nerlove’s results? Nerlove’s estimate of β_2 , for example, is 0.721 with a standard error of 0.0174 (the standard error in his paper is 0.175 but it’s probably a typo). Where in Nerlove’s paper can you find this estimate? What about the other coefficients? (Warning: You will not be able to replicate Nerlove’s results precisely. One reason is that he used common rather than natural logarithms; however, this should affect only the estimated intercept term. The other reason: the data set used for his results is a corrected version of the data set published with his article.)

As mentioned in the text, the plot of residuals suggests a nonlinear relationship between $\log(TC)$ and $\log(Q)$. Nerlove hypothesized that estimated returns to scale varied with the level of output. Following Nerlove, divide the sample

of 145 firms into five sub-sample or groups, each having 29 firms. (Recall that since the data are ordered by level of output, the first 29 observations will have the smallest output levels, whereas the last 29 observations will have the largest output levels.) Consider the following three generalizations of the model (1.7.6):

Model 1: Both the coefficients (β 's) and the error variance in (1.7.6) differ across groups.

Model 2: The coefficients are different but the error variance is the same across groups.

Model 3: While each group has common coefficients for β_3 and β_4 (price elasticities) and common error variance, it has a different intercept term and a different β_2 . Model 3 is what Nerlove called the hypothesis of neutral variations in returns to scale.

For Model 1, the coefficients and error variances specific to groups can be estimated from

$$\mathbf{y}^{(j)} = \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)} + \boldsymbol{\varepsilon}^{(j)} \quad (j = 1, \dots, 5),$$

where $\mathbf{y}^{(j)}$ (29×1) is the vector of the values of the dependent variable for group j , $\mathbf{X}^{(j)}$ (29×4) is the matrix of the values of the four regressors for group j , $\boldsymbol{\beta}^{(j)}$ (4×1) is the coefficient vector for group j , and $\boldsymbol{\varepsilon}^{(j)}$ (29×1) is the error vector. The second column of $\mathbf{X}^{(5)}$, for example, is $\log(Q)$ for $i = 117, \dots, 145$. Model 1 assumes conditional homoskedasticity $E(\boldsymbol{\varepsilon}^{(j)}\boldsymbol{\varepsilon}^{(j)'} | \mathbf{X}^{(j)}) = \sigma_j^2 \mathbf{I}_{29}$ within (but not necessarily across) groups.

- (d) Estimate Model 1 by OLS. How well can you replicate Nerlove's reported results? On the basis of your estimates of β_2 , compute the point estimates of returns to scale in each of the five groups. What is the general pattern of estimated scale economies as the level of output increases? What is the general pattern of the estimated error variance as output increases?

Model 2 assumes for Model 1 that $\sigma_j^2 = \sigma^2$ for all j . This equi-variance restriction can be incorporated by stacking vectors and matrices as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\underset{(145 \times 1)}{\mathbf{y}} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \vdots \\ \mathbf{y}^{(5)} \end{bmatrix}, \quad \underset{(145 \times 20)}{\mathbf{X}} = \begin{bmatrix} \mathbf{X}^{(1)} & & \\ & \ddots & \\ & & \mathbf{X}^{(5)} \end{bmatrix}, \quad \underset{(145 \times 1)}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \boldsymbol{\varepsilon}^{(1)} \\ \vdots \\ \boldsymbol{\varepsilon}^{(5)} \end{bmatrix}. \quad (*)$$

In particular, \mathbf{X} is now a block-diagonal matrix. The equi-variance restriction can be expressed as $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}) = \sigma^2\mathbf{I}_{145}$. There are now 20 variables derived from the original four regressors. The 145 dimensional vector corresponding to the second variable, for example, has $\log(Q_1), \dots, \log(Q_{29})$ as the first 29 elements and zeros elsewhere. The vector corresponding to the 6th variable, which represents log output for the second group of firms, has $\log(Q_{30}), \dots, \log(Q_{58})$ for the 30th through 58th elements and zeros elsewhere, and so on.

The stacking operation needed to form the \mathbf{y} and \mathbf{X} in (*) can be done easily if your computer software is matrix-based. Otherwise, you trick your software into accomplishing the same thing by the use of **dummy variables**. Define the j -th dummy variable as

$$D_{ji} = \begin{cases} 1 & \text{if firm } i \text{ belongs to the } j\text{-th group,} \\ 0 & \text{otherwise,} \end{cases} \quad (i = 1, \dots, 145).$$

Then the second regressor is $D_{1i} \cdot \log(Q_i)$. The 6th variable is $D_{2i} \cdot \log(Q_i)$, and so forth.

- (e) Estimate Model 2 by OLS. Verify that the OLS coefficient estimates here are the same as those in (d). Also verify that

$$\sum_{j=1}^5 SSR_j = SSR,$$

where SSR_j is the SSR from the j -th group in your estimation of Model 1 in (d) and SSR is the SSR from Model 2. This agreement is not by accident, i.e., not specific to the present data set. Prove that this agreement for the coefficients and the SSR holds in general, temporarily assuming just two groups without loss of generality. **Hint:** First show that the coefficient estimate is the same between Model 1 and Model 2. Use formulas (A.4), (A.5), and (A.9) of Appendix A.

- (f) (Chow test) Model 2 is more general than the Model (1.7.6) because the coefficients can differ across groups. Test the null hypothesis that the coefficients are the same across groups. How many equations (restrictions) in the null hypothesis? This test is sometimes called the **Chow test for structural change**. Calculate the p -value of the F -ratio. **Hint:** This is a linear hypothesis about the coefficients of Model 2. So take Model 2 to be the maintained hypothesis and (1.7.6) to be the restricted model. Use the formula (1.4.11) for the F -ratio.

Gauss Tip: If x is the F -ratio, the Gauss command `cdfFc(x, df1, df2)` gives the area to the right of F for the F distribution with $df1$ and $df2$ degrees of freedom.

TSP Tip: The TSP command to do the same is: `cdf (f, df1=df1, df2=df2) x`. An output of TSP's OLS command, `OLSQ`, is `@SSR` which is the *SSR* for the regression.

RATS Tip: The RATS command is: `cdf ftest x df1 df2`. An output of RATS's OLS command, `LINREG`, is `%RSS` which is the *SSR* for the regression.

The restriction in Model 3 that the price elasticities are the same across firm groups can be imposed on Model 2 by applying the dummy variable transformation only to the constant and log output. Thus there are 12 ($= 2 \times 5 + 2$) variables in \mathbf{X} . Now \mathbf{X} looks like:

$$\mathbf{X} = \begin{bmatrix} 1 & \log(Q_1) & 0 & 0 & \log(PL_1/PF_1) & \log(PK_1/PF_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \log(Q_{29}) & 0 & 0 & \log(PL_{29}/PF_{29}) & \log(PK_{29}/PF_{29}) \\ & & \ddots & & \vdots & \vdots \\ 0 & 0 & 1 & \log(Q_{117}) & \log(PL_{117}/PF_{117}) & \log(PK_{117}/PF_{117}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & \log(Q_{145}) & \log(PL_{145}/PF_{145}) & \log(PK_{145}/PF_{145}) \end{bmatrix} \quad (**)$$

- (g) Estimate Model 3. The model is a special case of Model 2, with the hypothesis that the two price elasticities are the same across the five groups. Test the hypothesis at a significance level of 5%, assuming normality. (Note: Nerlove's *F*-ratio on p. 183 is wrong.)

As has become clear from the plot of residuals in Figure 1.7, the conditional second moment $E(\varepsilon_i^2 | \mathbf{X})$ is likely to depend on log output, which is a violation of the conditional homoskedasticity assumption. This time we don't attempt to test conditional homoskedasticity, because to do so requires large sample theory and is postponed until the next chapter. Instead, we pretend to know the form of the function linking the conditional second moment to log output. The function, specified below, implies that the conditional second moment varies continuously with output, contrary to the three models we have considered above. Also contrary to those models, we assume that the degree of returns to scale varies continuously with output by including the square of log output.²⁵ Model 4 is

²⁵We have derived the log-linear cost function from the Cobb-Douglas production function. Does there exist a production function from which this generalized cost function with a quadratic term in log output can be derived? This is a question of the "integrability" of cost functions, and discussed in detail in Christensen et al. (1973).

Model 4:

$$\log\left(\frac{TC_i}{p_{i3}}\right) = \beta_1 + \beta_2 \log(Q_i) + \beta_3 [\log(Q_i)]^2 + \beta_4 \log\left(\frac{p_{i1}}{p_{i3}}\right) + \beta_5 \log\left(\frac{p_{i2}}{p_{i3}}\right) + \varepsilon_i$$
$$E(\varepsilon_i^2 | \mathbf{X}) = \sigma^2 \cdot \left(0.0565 + \frac{2.1377}{Q_i}\right) \quad (i = 1, 2, \dots, 145)$$

for some unknown σ^2 .

- (h) Estimate Model 4 by weighted least squares on the whole sample of 145 firms. (Be careful about the treatment of the intercept; in the equation after weighting, none of the regressors is a constant.) Plot the residuals. Is there still evidence for conditional homoskedasticity or further non-linearities?

Monte Carlo Exercise

Monte Carlo analysis simulates a large number of samples from the model to study the finite-sample distribution of estimators. In this exercise, we use the technique to confirm the two finite-sample results of the text: the unbiasedness of the OLS coefficient estimator and the distribution of the t -ratio. The model is the following simple regression model satisfying Assumptions 1.1–1.5 with $n = 32$. The regression equation is

$$\begin{aligned} y_i &= \beta_1 + \beta_2 x_i + \varepsilon_i \quad (i = 1, 2, \dots, n) \\ \text{or } \mathbf{y} &= \mathbf{1} \cdot \beta_1 + \mathbf{x} \cdot \beta_2 + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \end{aligned} \quad (*)$$

where $\mathbf{X} = (\mathbf{1} : \mathbf{x})$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)'$. The model parameters are $(\beta_1, \beta_2, \sigma^2)$.

As mentioned in the text, a model is a set of joint distributions of (\mathbf{y}, \mathbf{X}) . We pick a particular joint distribution by specifying the regression model as follows. Set $\beta_1 = 1$, $\beta_2 = 0.5$, and $\sigma^2 = 1$. The distribution of $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ is specified by the following AR(1) process:

$$x_i = c + \phi x_{i-1} + \eta_i \quad (i = 1, 2, \dots, n), \quad (**)$$

where $\{\eta_i\}$ is i.i.d. $N(0, 1)$ and

$$x_0 \sim N\left(\frac{c}{1-\phi}, \frac{1}{1-\phi^2}\right), \quad c = 2, \quad \phi = 0.6.$$

This fixes the joint distribution of (\mathbf{y}, \mathbf{X}) . From this distribution, a large number of samples will be drawn.

In programming the simulation, the following expression for \mathbf{x} will be useful. Solve the first-order difference equation (**) to obtain

$$x_i = \phi^i x_0 + (1 + \phi + \phi^2 + \dots + \phi^{i-1})c + (\eta_i + \phi\eta_{i-1} + \phi^2\eta_{i-2} + \dots + \phi^{i-1}\eta_1),$$

or, in matrix notation,

$$\underset{(n \times 1)}{\mathbf{x}} = \underset{(n \times 1)}{\mathbf{r}} \cdot x_0 + \underset{(n \times 1)}{\mathbf{d}} + \underset{(n \times n)}{\mathbf{A}} \underset{(n \times 1)}{\boldsymbol{\eta}} \quad (***)$$

where $\mathbf{d} = (d_1, d_2, \dots, d_n)'$ and

$$\begin{aligned} d_1 &= c, \quad d_2 = (1 + \phi)c, \dots, \quad d_i = (1 + \phi + \phi^2 + \dots + \phi^{i-1})c, \dots, \\ \mathbf{r} &= \begin{bmatrix} \phi \\ \phi^2 \\ \vdots \\ \phi^n \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \phi & 1 & 0 & \dots & 0 \\ \phi^2 & \phi & 1 & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \dots & \phi & 1 \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}. \end{aligned}$$

Gauss Tip: To form the \mathbf{r} matrix, use `seqm`. To form the \mathbf{A} matrix, use `toeplitz` and `lowmat`.

(a) Run two Monte Carlo simulations. The first simulation calculates $E(\mathbf{b} \mid \mathbf{x})$ and the distribution of the t -ratio as a distribution conditional on \mathbf{x} . A computer program for the first simulation should consist of the following steps.

- (1) (generate \mathbf{x} just once) Using the random number generator, draw a vector $\boldsymbol{\eta}$ of n i.i.d. random variables from $N(0, 1)$ and x_0 from $N(c/(1 - \phi), 1/(1 - \phi^2))$, and calculate \mathbf{x} by `(***)`. (Calculation of \mathbf{x} can also be accomplished recursively by `(**)` with a do loop, but vector operations such as `(***)` consume less CPU time than do loops. This becomes a consideration in the second simulation, where \mathbf{x} has to be generated in each replication.)
- (2) Set a counter to zero. The counter will record the incidence that $|t| > t_{0.025}(n - 2)$. Also, set a two-dimensional vector at zero; this vector will be used for calculating the mean of the OLS estimator \mathbf{b} of $(\beta_1, \beta_2)'$.
- (3) Start a do loop of a large number of replications (1 million, say). In each replication, do the following.
 - (i) (generate \mathbf{y}) Draw an n dimensional vector $\boldsymbol{\varepsilon}$ of n i.i.d. random variables from $N(0, 1)$, and calculate $\mathbf{y} = (y_1, \dots, y_n)'$ by `(*)`. This \mathbf{y} is paired with the same \mathbf{x} from step (1) to form a sample (\mathbf{y}, \mathbf{x}) .
 - (ii) From the sample, calculate the OLS estimator \mathbf{b} and the t -value for $H_0: \beta_2 = 0.5$.
 - (iii) Increase the counter by one if $|t| > t_{0.025}(n - 2)$. Also, add \mathbf{b} to the two-dimensional vector.
- (4) After the do loop, divide the counter by the number of replications to calculate the frequency of rejecting the null. Also, divide the two-dimensional vector that has accumulated \mathbf{b} by the number of replications. It should equal $E(\mathbf{b} \mid \mathbf{x})$ if the number of replications is infinite.

Note that in this first simulation, \mathbf{x} is *fixed* throughout the do loop for \mathbf{y} . The second simulation calculates the *unconditional* distribution of the t -ratio. It should consist of the following steps.

- (1) Set the counter to zero.
- (2) Start a do loop of a large number of replications. In each replication, do the following.
 - (i) (generate \mathbf{x}) Draw a vector $\boldsymbol{\eta}$ of n i.i.d. random variables from $N(0, 1)$ and x_0 from $N(c/(1 - \phi), 1/(1 - \phi^2))$, and calculate \mathbf{x} by `(***)`.

- (ii) (generate \mathbf{y}) Draw a vector $\boldsymbol{\varepsilon}$ of n i.i.d. random variables from $N(0, 1)$, and calculate $\mathbf{y} = (y_1, \dots, y_n)'$ by (*).
 - (iii) From a sample (\mathbf{y}, \mathbf{x}) thus generated, calculate the t -value for $H_0: \beta = 0.5$ from the sample (\mathbf{y}, \mathbf{x}) .
 - (iv) Increase the counter by one if $|t| > t_{0.025}(n - 2)$.
- (3) After the do loop, divide the counter by the number of replications.

For the two simulations, verify that, for a sufficiently large number of replications,

1. the mean of \mathbf{b} from the first simulation is arbitrarily close to the true value $(1, 0.5)$;
 2. the frequency of rejecting the true hypothesis H_0 (the type I error) is arbitrarily close to 5% in either simulation.
- (b) In those two simulations, is the (non-constant) regressor strictly exogenous? Is the error conditionally homoskedastic?

Answers to Selected Questions

Analytical Exercises

1.

$$\begin{aligned}
& (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\
&= [(\mathbf{y} - \mathbf{X}\mathbf{b}) + \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}})]'[(\mathbf{y} - \mathbf{X}\mathbf{b}) + \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}})] \\
&\quad \text{(by the "add-and-subtract" strategy)} \\
&= [(\mathbf{y} - \mathbf{X}\mathbf{b})' + (\mathbf{b} - \tilde{\boldsymbol{\beta}})' \mathbf{X}'][(\mathbf{y} - \mathbf{X}\mathbf{b}) + \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}})] \\
&= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \tilde{\boldsymbol{\beta}})' \mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\
&\quad + (\mathbf{y} - \mathbf{X}\mathbf{b})' \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}) + (\mathbf{b} - \tilde{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}) \\
&= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + 2(\mathbf{b} - \tilde{\boldsymbol{\beta}})' \mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \tilde{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}) \\
&\quad \text{(since } (\mathbf{b} - \tilde{\boldsymbol{\beta}})' \mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})' \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}})) \\
&= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \tilde{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}) \\
&\quad \text{(since } \mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{0} \text{ by the normal equations)} \\
&\geq (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\
&\quad \text{(since } (\mathbf{b} - \tilde{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}}) = \mathbf{z}'\mathbf{z} = \sum_{i=1}^n z_i^2 \geq 0 \text{ where } \mathbf{z} \equiv \mathbf{X}(\mathbf{b} - \tilde{\boldsymbol{\beta}})).
\end{aligned}$$

7(a) $\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta} = \mathbf{A}\boldsymbol{\varepsilon}$ where $\mathbf{A} \equiv (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$ and $\mathbf{b} - \hat{\boldsymbol{\beta}}_{\text{GLS}} = \mathbf{B}\boldsymbol{\varepsilon}$ where $\mathbf{B} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. So

$$\begin{aligned}
& \text{Cov}(\hat{\boldsymbol{\beta}}_{\text{GLS}} - \boldsymbol{\beta}, \mathbf{b} - \hat{\boldsymbol{\beta}}_{\text{GLS}}) \\
&= \text{Cov}(\mathbf{A}\boldsymbol{\varepsilon}, \mathbf{B}\boldsymbol{\varepsilon}) \\
&= \mathbf{A} \text{Var}(\boldsymbol{\varepsilon})\mathbf{B}' \\
&= \sigma^2 \mathbf{A}\mathbf{V}\mathbf{B}'.
\end{aligned}$$

It is straightforward to show that $\mathbf{A}\mathbf{V}\mathbf{B}' = \mathbf{0}$.

7(b) For the choice of \mathbf{H} indicated in the hint,

$$\text{Var}(\hat{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}}) = -\mathbf{C}\mathbf{V}_q^{-1}\mathbf{C}'.$$

If $\mathbf{C} \neq \mathbf{0}$, then there exists a non-zero vector \mathbf{z} such that $\mathbf{C}'\mathbf{z} \equiv \mathbf{v} \neq \mathbf{0}$. For such \mathbf{z} ,

$$\mathbf{z}'[\text{Var}(\hat{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}}_{\text{GLS}})]\mathbf{z} = -\mathbf{v}'\mathbf{V}_q^{-1}\mathbf{v} < 0 \quad \text{(since } \mathbf{V}_q \text{ is positive definite),}$$

which is a contradiction because $\hat{\boldsymbol{\beta}}_{\text{GLS}}$ is efficient.

Empirical Exercise

- (a) Nerlove's description in Appendix B.4 leads one to believe that he didn't include the depreciation rate δ in his construction of the price of capital.
- (b) Your estimates should agree with (1.7.7).
- (c) Our estimates differ from Nerlove's slightly. This would happen even if the data used by Nerlove were the same as those provided to you because computers in his age had much less precision with more frequent rounding errors.
- (d) How well can you replicate Nerlove's reported results? Fairly well. The point estimates of returns to scale in each of the five sub-samples are: 2.5, 1.5, 1.1, 1.1, and .96. As the level of output increases, the returns to scale decline.
- (e) Model 2 can be written as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where \mathbf{y} , \mathbf{X} , and $\boldsymbol{\varepsilon}$ are as in (*). So (setting $j = 2$),

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)'}\mathbf{X}^{(1)} & \mathbf{0} \\ \mathbf{0} & \mathbf{X}^{(2)'}\mathbf{X}^{(2)} \end{bmatrix},$$

which means

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} (\mathbf{X}^{(1)'}\mathbf{X}^{(1)})^{-1} & \mathbf{0} \\ \mathbf{0} & (\mathbf{X}^{(2)'}\mathbf{X}^{(2)})^{-1} \end{bmatrix}.$$

And

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \mathbf{X}^{(1)'}\mathbf{y}^{(1)} \\ \mathbf{X}^{(2)'}\mathbf{y}^{(2)} \end{bmatrix}.$$

Therefore,

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} (\mathbf{X}^{(1)'}\mathbf{X}^{(1)})^{-1}\mathbf{X}^{(1)'}\mathbf{y}^{(1)} \\ (\mathbf{X}^{(2)'}\mathbf{X}^{(2)})^{-1}\mathbf{X}^{(2)'}\mathbf{y}^{(2)} \end{bmatrix}.$$

Thus, the OLS estimate of the coefficient vector for Model 2 is the same as that for Model 1. Since the estimate of the coefficient vector is the same, the sum of squared residuals, too, is the same.

- (f) The number of restrictions is 16. $K = \#$ coefficients in Model 2 = 20. So the two degrees of freedom should be (16, 125). $SSR_U = 12.262$ and $SSR_R = 21.640$. F -ratio = 5.97 with a p -value of 0.0000. So can reject at any reasonable significance level.

- (g) $SSR_U = 12.262$ and $SSR_R = 12.577$. So $F = .40$ with 8 and 125 degrees of freedom. Its p -value is 0.92. So the restrictions can be accepted at any reasonable significance level. Nerlove's F -ratio (see p. 183, 8th line from bottom) is 1.576.
- (h) The plot still shows that the conditional second moment is somewhat larger for smaller firms, but now there is no evidence for possible nonlinearities.

References

- Amemiya, T., 1985, *Advanced Econometrics*, Harvard University Press, Cambridge.
- Averch, H., and L. Johnson, 1962, "Behavior of the Firm under Regulatory Constraint," *American Economic Review*, 52, 1052-1069.
- Christensen, L., and W. Greene, 1976, "Economies of Scale in US Electric Power Generation," *Journal of Political Economy*, 84, 655-676.
- Christensen, L., D. Jorgenson, and L. Lau, 1973, "Transcendental Logarithmic Production Frontiers," *Review of Economics and Statistics*, 55, 28-45.
- Davidson, R., and J. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford University Press.
- DeLong, B., and L. Summers, 1991, "Equipment Investment and Growth," *Quarterly Journal of Economics*, 99, 28-45.
- Engle, R., D. Hendry, and J.-F. Richards, 1983, "Exogeneity," *Econometrica*, 51, 277-304.
- Federal Power Commission, 1956, *Statistics of Electric Utilities in the United States, 1955, Class A and B Privately Owned Companies*, Washington, D.C.
- Jorgenson, D., 1963, "Capital Theory and Investment Behavior," *American Economic Review*, 53, 247-259.
- Koopmans, T., and W. Hood, 1953, "The Estimation of Simultaneous Linear Economic Relationships," in W. Hood, and T. Koopmans (eds.), *Studies in Econometric Method*, Yale University Press, New Haven.
- Krasker, W., E. Kuh, and R. Welsch, 1983, "Estimation for Dirty Data and Flawed Models," in Z. Griliches, and M. Intriligator (eds.), *Handbook of Econometrics* (vol. 1), chap. 11, North-Holland.
- Nerlove, M., 1963, "Returns to Scale in Electricity Supply," in C. Christ (ed.), *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford University Press.
- Rao, C. R., 1973, *Linear Statistical Inference and Its Applications* (2nd ed.), John Wiley & Sons, New York.
- Scheffe, H., 1959, *The Analysis of Variance*, John Wiley & Sons, New York.

- Wolak, F., 1994, "An Econometric Analysis of the Asymmetric Information, Regulator-Utility Interaction," *Annales D'Economie et de Statistique*, 34, 13-69.

Figure 1.1: Hypothetical, True, and Estimated Values

Figure 1.2: Equipment Investment and Growth

Figure 1.3: t Distribution

Figure 1.4: F Distribution

Figure 1.5: t vs. F Tests

Figure 1.6: Output Determination

Figure 1.7: Plot of Residuals against Output

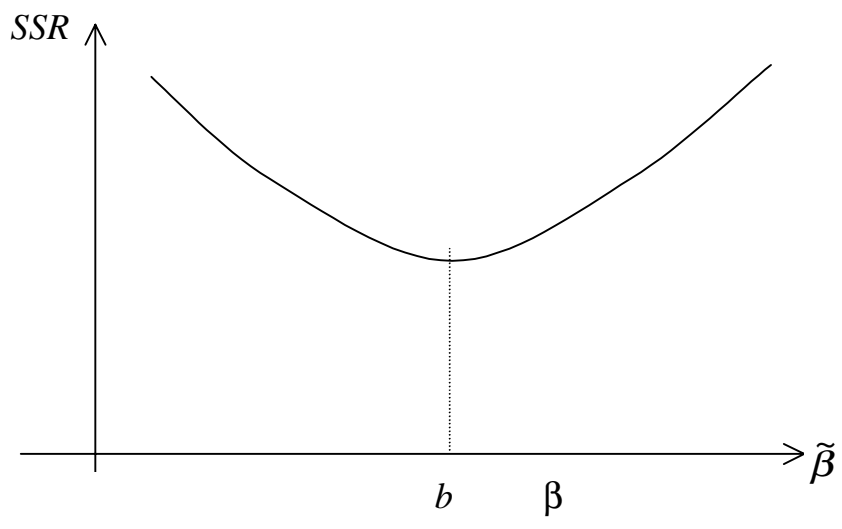


Figure 1.1

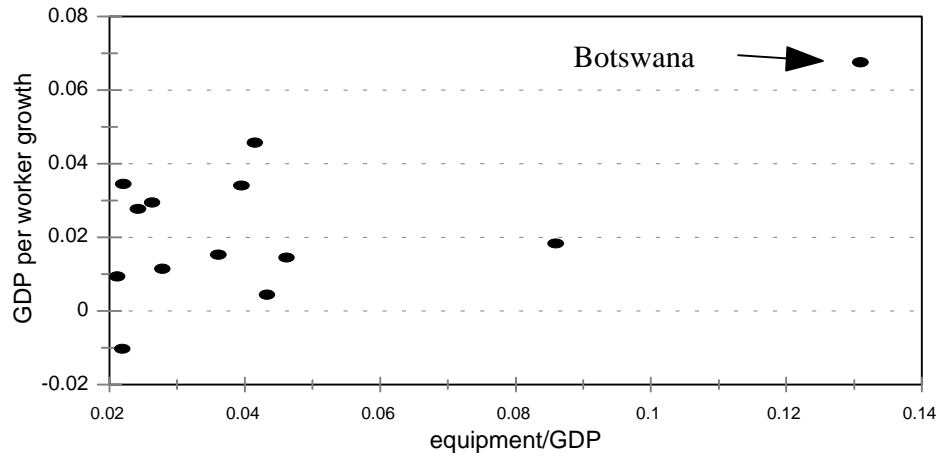


Figure 1.2
equipment investment and growth

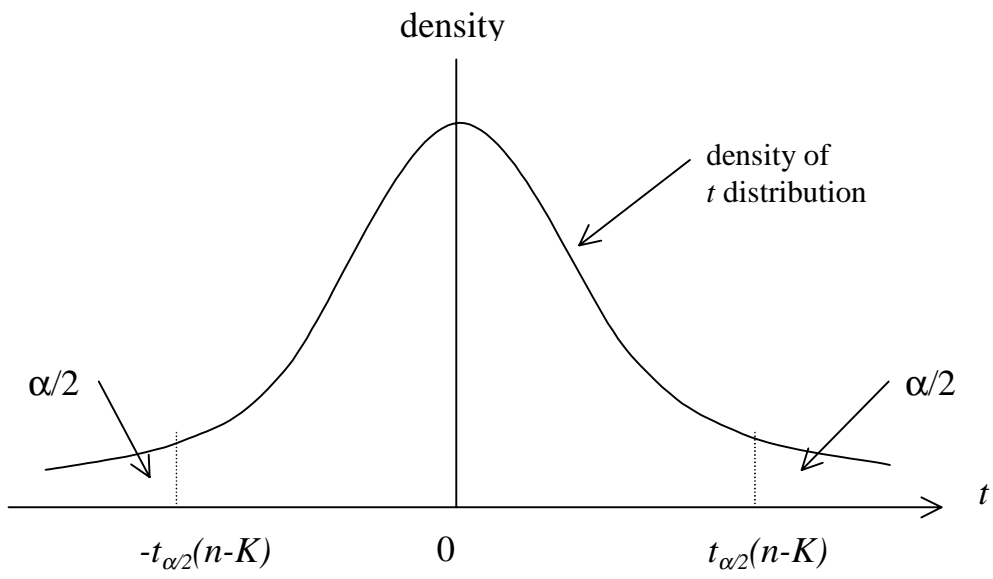


Figure 1.3

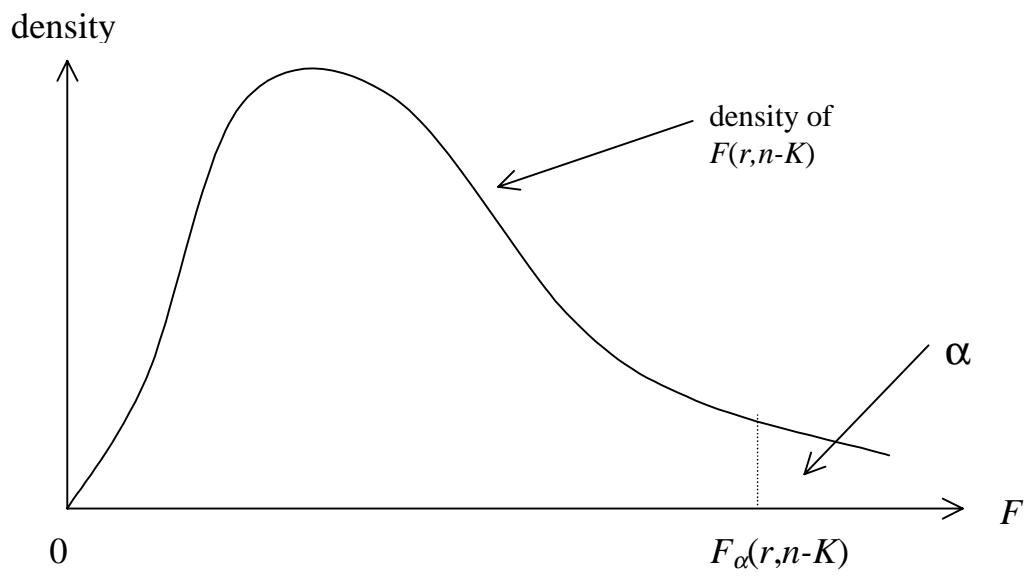


Figure 1.4

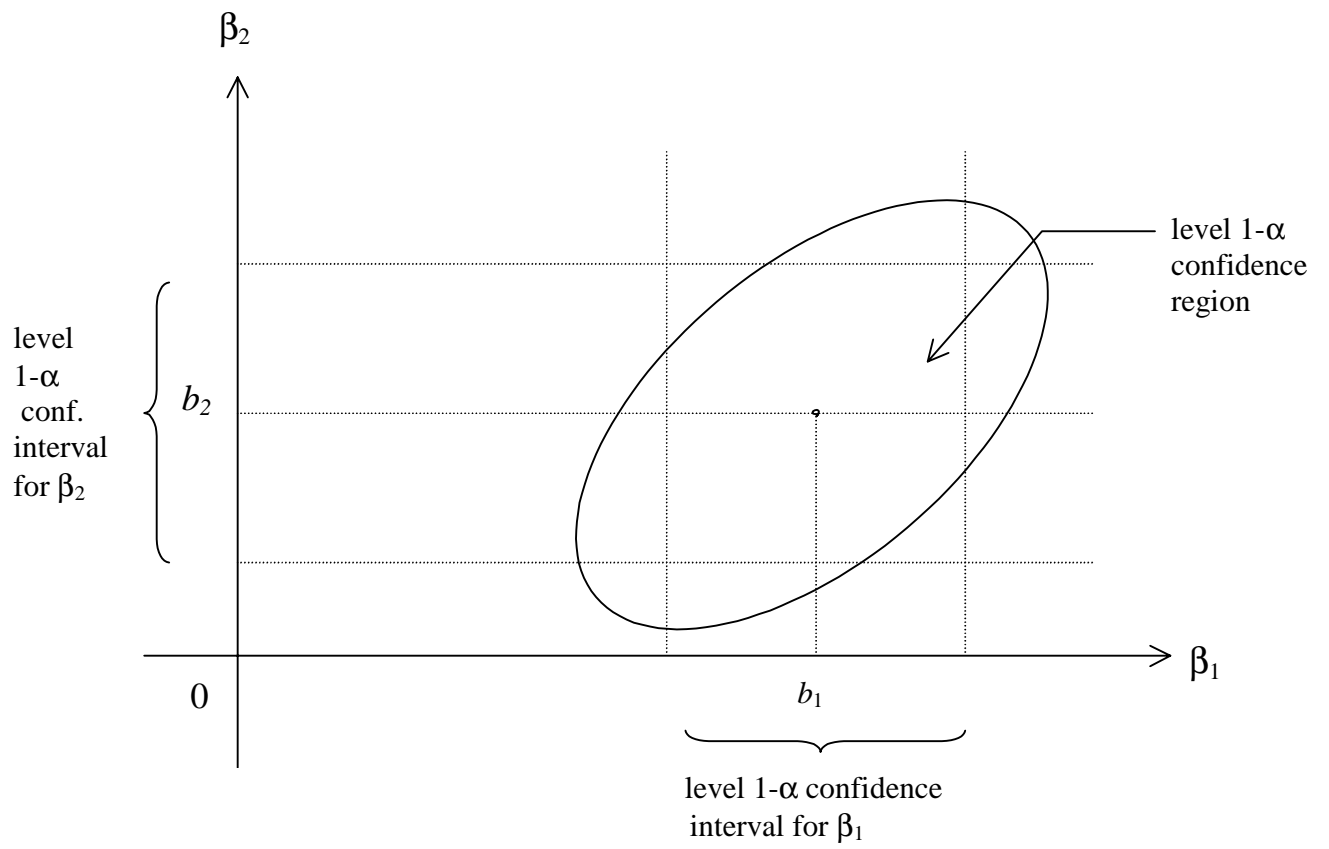


Figure 1.5

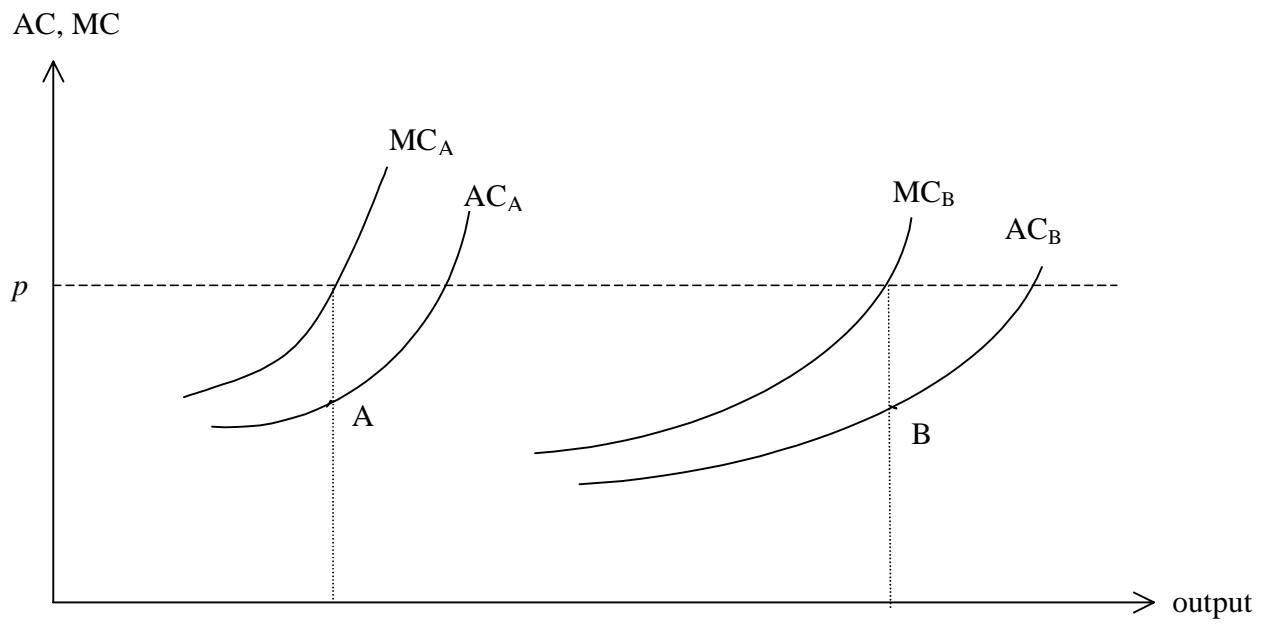


Figure 1.6

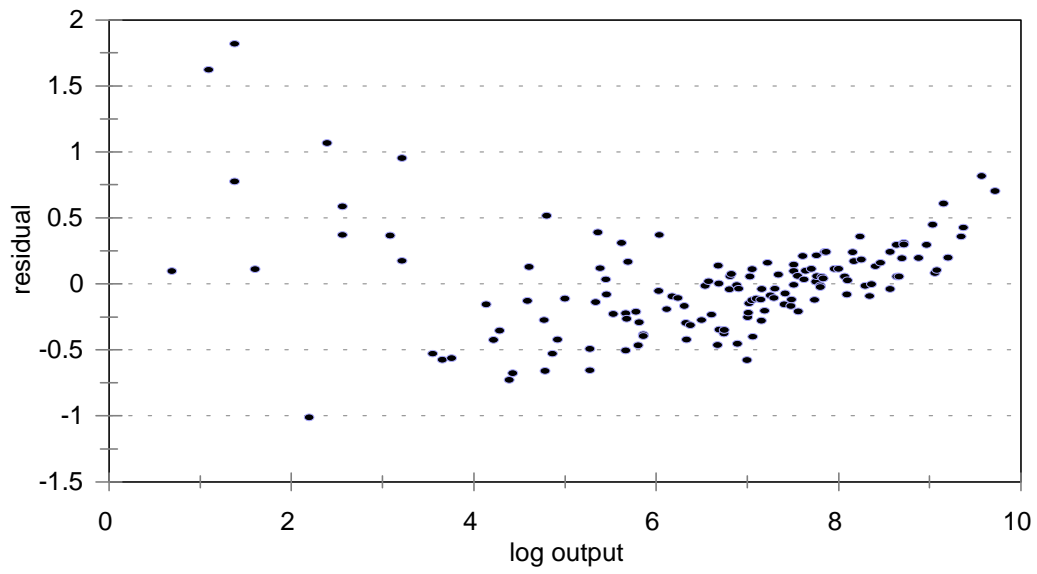


Figure 1.7
plot of residuals against log output

1.5 Relation to Maximum Likelihood

Having specified the distribution of the error vector $\boldsymbol{\varepsilon}$, we can use the **maximum likelihood (ML) principle** to estimate the model parameters $(\boldsymbol{\beta}, \sigma^2)$.²¹ In this section, we will show that \mathbf{b} , the OLS estimator of $\boldsymbol{\beta}$, is also the ML estimator, and the OLS estimator of σ^2 differs only slightly from the ML counterpart, when the error is normally distributed. We will also show that \mathbf{b} achieves the **Cramer-Rao lower bound**.

The Maximum Likelihood Principle

Just to refresh your memory of basic statistics, we temporarily step outside the classical regression model to review the ML estimation and related concepts. The basic idea of the ML principle is to choose the parameter estimates to maximize the probability of obtaining the data. To be more precise, suppose that we observe an n -dimensional data vector $\mathbf{y} \equiv (y_1, y_2, \dots, y_n)'$. Assume that the probability density of \mathbf{y} is a member of a family of functions indexed by a finite-dimensional parameter vector $\boldsymbol{\theta}$: $f(\mathbf{y}; \boldsymbol{\theta})$. The set of values that $\boldsymbol{\theta}$ could take is called the **parameter space** and denoted by Θ . (This is described as **parameterizing** the density function.) When the hypothetical parameter vector $\tilde{\boldsymbol{\theta}}$ equals the true parameter vector $\boldsymbol{\theta}$, $f(\mathbf{y}; \tilde{\boldsymbol{\theta}})$ becomes the true density of \mathbf{y} . We have thus specified a model, a set of possible distributions of \mathbf{y} . The model is said to be **correctly specified** if the parameter space Θ includes the true parameter value $\boldsymbol{\theta}$.

The hypothetical density $f(\mathbf{y}; \tilde{\boldsymbol{\theta}})$, viewed as a function of the hypothetical parameter vector $\tilde{\boldsymbol{\theta}}$, is called the **likelihood function** $L(\tilde{\boldsymbol{\theta}})$. Thus,

$$L(\tilde{\boldsymbol{\theta}}) \equiv f(\mathbf{y}; \tilde{\boldsymbol{\theta}}). \quad (1.5.1)$$

The ML estimate of the unknown true parameter vector $\boldsymbol{\theta}$ is the $\tilde{\boldsymbol{\theta}}$ that maximizes the likelihood function. The maximization is equivalent to maximizing the **log likelihood function** $\log L(\tilde{\boldsymbol{\theta}})$ because the log transformation is a monotone transformation. Therefore, the ML estimator of $\boldsymbol{\theta}$ can be defined as

$$\text{ML estimator of } \boldsymbol{\theta} \equiv \underset{\tilde{\boldsymbol{\theta}} \in \Theta}{\operatorname{argmax}} \log L(\tilde{\boldsymbol{\theta}}). \quad (1.5.2)$$

For example, consider the so-called normal location/scale model:

Example 1.6: Suppose the sample \mathbf{y} is an i.i.d. sample from $N(\mu, \sigma^2)$ (the normal distribution with mean μ and variance σ^2). With $\boldsymbol{\theta} \equiv (\mu, \sigma^2)'$, the (joint) density of \mathbf{y} is given by

$$f(\mathbf{y}; \boldsymbol{\theta}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(y_i - \mu)^2}{2\sigma^2} \right]. \quad (1.5.3)$$

We take the parameter space Θ to be $\Re \times \Re_{++}$, where \Re_{++} is a set of positive real numbers. This just means that μ can be any real number but σ^2 is constrained

²¹For a fuller treatment of maximum likelihood, see Chapter 7.

to be positive. Replacing the true parameter value $\boldsymbol{\theta}$ by its hypothetical value $\tilde{\boldsymbol{\theta}} \equiv (\tilde{\mu}, \tilde{\sigma}^2)'$ and then taking logs, we obtain the log likelihood function:

$$\log L(\tilde{\boldsymbol{\theta}}) = -\frac{n}{2} \log(2\pi\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2} \sum_{i=1}^n (y_i - \tilde{\mu})^2. \quad (1.5.4)$$

The Score, the Information Matrix, and the Cramer-Rao Bound

Having introduced the log likelihood function, we can now define some concepts related to the ML estimation. The **score function** or the **score** is simply the gradient (vector of partial derivatives) of log likelihood:

$$\text{score: } \mathbf{s}(\tilde{\boldsymbol{\theta}}) \equiv \frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}}. \quad (1.5.5)$$

The **information matrix**, denoted as $\mathbf{I}(\boldsymbol{\theta})$, is defined as the matrix of second moments of the score *evaluated at the true parameter vector* $\boldsymbol{\theta}$:

$$\mathbf{I}(\boldsymbol{\theta}) \equiv \text{E}[\mathbf{s}(\boldsymbol{\theta}) \mathbf{s}(\boldsymbol{\theta})']. \quad (1.5.6)$$

The celebrated Cramer-Rao inequality states that the variance of any unbiased estimator is greater than or equal to the inverse of the information matrix.

Cramer-Rao Inequality: *Let \mathbf{y} be a vector of random variables (not necessarily independent) the joint density of which is given by $f(\mathbf{y}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is an K -dimensional vector of parameters. Let $L(\tilde{\boldsymbol{\theta}}) \equiv f(\mathbf{y}; \tilde{\boldsymbol{\theta}})$ be the likelihood function, and let $\hat{\boldsymbol{\theta}}(\mathbf{y})$ be an unbiased estimator of $\boldsymbol{\theta}$ with a finite variance-covariance matrix. Then, under some regularity conditions on $f(\mathbf{y}; \boldsymbol{\theta})$ (not stated here),*

$$\text{Var}[\hat{\boldsymbol{\theta}}(\mathbf{y})] \geq \mathbf{I}(\boldsymbol{\theta})^{-1} \quad (\equiv \text{Cramer-Rao Lower Bound}),$$

$(K \times K)$

Also under the regularity conditions, the information matrix equals the negative of the expected value of the Hessian (matrix of second partial derivatives) of the log likelihood evaluated at the true parameter vector $\boldsymbol{\theta}$:

$$\mathbf{I}(\boldsymbol{\theta}) = -\text{E} \left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\theta}}'} \right]. \quad (1.5.7)$$

*This is called the **information matrix equality**.*

See, e.g., Amemiya (1985, Theorem 1.3.1) for a proof and a statement of the regularity conditions. Those conditions guarantee that the operations of differentiation and taking expectations can be interchanged. Thus, for example,

$$\text{E}[\partial L(\boldsymbol{\theta}) / \partial \tilde{\boldsymbol{\theta}}] = \partial \text{E}[L(\boldsymbol{\theta})] / \partial \tilde{\boldsymbol{\theta}}.$$

Conditional Likelihood

The theory of maximum likelihood presented above is based on the hypothetical density $f(\mathbf{y}; \tilde{\boldsymbol{\theta}})$. You may have noticed that all that is required for the above results to hold is that $f(\mathbf{y}; \tilde{\boldsymbol{\theta}})$ is a density, so the whole discussion can be easily adapted to a conditional density. That is, if the data are (\mathbf{y}, \mathbf{X}) rather than \mathbf{y} , we can develop the conditional version of the theory of ML and related concepts, based on the hypothetical conditional density $f(\mathbf{y} | \mathbf{X}; \tilde{\boldsymbol{\theta}})$, as follows.

- The likelihood function $L(\tilde{\boldsymbol{\theta}})$ is now the conditional likelihood

$$L(\tilde{\boldsymbol{\theta}}) \equiv f(\mathbf{y} \mid \mathbf{X}; \tilde{\boldsymbol{\theta}}). \quad (1.5.8)$$

- Just as above, the ML estimator is the $\tilde{\boldsymbol{\theta}}$ that maximizes this likelihood function. The estimator is called the **conditional ML estimator**. The score function is defined just as above, as the gradient of the log likelihood function.
- The definition of the information matrix is the same as above, except that the expectation is conditional on \mathbf{X} :

$$\mathbf{I}(\boldsymbol{\theta}) \equiv \text{E}[\mathbf{s}(\boldsymbol{\theta}) \mathbf{s}(\boldsymbol{\theta})' \mid \mathbf{X}]. \quad (1.5.9)$$

The inverse of this matrix is the Cramer-Rao lower bound.

- Similarly, the expectation in the information matrix equality is conditional on \mathbf{X} :

$$\mathbf{I}(\boldsymbol{\theta}) = -\text{E}\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\theta}}} \mid \mathbf{X}\right]. \quad (1.5.10)$$

Given the data (\mathbf{y}, \mathbf{X}) , we could also construct the theory based on the joint density of (\mathbf{y}, \mathbf{X}) . A natural question that arises is: what is the relationship between the conditional ML estimator, which is based on the density of \mathbf{y} given \mathbf{X} , and the (full or joint) ML estimator based on the joint density of (\mathbf{y}, \mathbf{X}) ? This issue will be briefly discussed at the end of this section and more fully in Chapter 7.

Conditional ML Estimation of the Classical Linear Regression Model

We now return to the classical regression model and derive the conditional ML estimator and the Cramer-Rao lower bound.

The Log Likelihood

As already observed, Assumption 1.5 (the normality assumption) together with Assumptions 1.2 and 1.4 imply that the distribution of $\boldsymbol{\varepsilon}$ conditional on \mathbf{X} is $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ (see (1.4.1)). But since $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ by Assumption 1.1, we have

$$\mathbf{y} \mid \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n). \quad (1.5.11)$$

Thus, the conditional density of \mathbf{y} given \mathbf{X} is²²

$$f(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]. \quad (1.5.12)$$

²²Recall from basic probability theory that the density function for an n -variate normal distribution with mean $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ is

$$(2\pi)^{-n/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\mu})\right].$$

To derive (1.5.12), just set $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_n$.

Replacing the true parameters $(\boldsymbol{\beta}, \sigma^2)$ by their hypothetical values $(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2)$ and taking logs, we obtain the log likelihood function:

$$\begin{aligned}\log L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2} SSR(\tilde{\boldsymbol{\beta}}),\end{aligned}\quad (1.5.13)$$

where $SSR(\tilde{\boldsymbol{\beta}}) \equiv (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$ is the sum of squared residuals considered in Section 1.2.

ML via Concentrated Likelihood

It is instructive to maximize the log likelihood in two stages. First, maximize over $\tilde{\boldsymbol{\beta}}$ for any given $\tilde{\sigma}^2$. The $\tilde{\boldsymbol{\beta}}$ that maximizes the objective function could (but does not, in the present case of Assumptions 1.1–1.5) depend on $\tilde{\sigma}^2$. Second, maximize over $\tilde{\sigma}^2$ taking into account that the $\tilde{\boldsymbol{\beta}}$ obtained in the first stage could depend on $\tilde{\sigma}^2$. The log likelihood function in which $\tilde{\boldsymbol{\beta}}$ is constrained to be the value from the first stage is called the **concentrated log likelihood function** (concentrated with respect to $\tilde{\boldsymbol{\beta}}$). For the normal log likelihood (1.5.13), the first stage amounts to minimizing the sum of squares $(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$. The $\tilde{\boldsymbol{\beta}}$ that does it is none other than the OLS estimator \mathbf{b} , and the minimized sum of squares is $\mathbf{e}'\mathbf{e}$. Thus, the concentrated log likelihood is

$$\text{concentrated log likelihood} = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2} \mathbf{e}'\mathbf{e}. \quad (1.5.14)$$

This is a function of $\tilde{\sigma}^2$ alone, and the $\tilde{\sigma}^2$ that maximizes the concentrated likelihood is the ML estimate of σ^2 . The maximization is straightforward for the present case of the classical regression model, because $\mathbf{e}'\mathbf{e}$ is not a function of $\tilde{\sigma}^2$ and so can be taken as a constant. Still, taking the derivative with respect to $\tilde{\sigma}^2$, rather than with respect to $\tilde{\sigma}$, can be tricky. This can be avoided by denoting $\tilde{\sigma}^2$ by $\tilde{\gamma}$. Taking the derivative of (1.5.14) with respect to $\tilde{\gamma}$ ($\equiv \tilde{\sigma}^2$) and setting it to zero, we obtain the following result.

Proposition 1.5 (ML Estimator of $(\boldsymbol{\beta}, \sigma^2)$): *Suppose Assumptions 1.1–1.5 hold. Then the ML estimator of $\boldsymbol{\beta}$ is the OLS estimator \mathbf{b} and*

$$\text{ML estimator of } \sigma^2 = \frac{1}{n} \mathbf{e}'\mathbf{e} = \frac{SSR}{n} = \frac{n-K}{n} s^2. \quad (1.5.15)$$

We know from Proposition 1.2 that s^2 is unbiased. Since s^2 is multiplied by a factor $(n-K)/n$ which is different from 1, the ML estimator of σ^2 is biased, although the bias becomes arbitrarily small as the sample size n increases for any given fixed K .

For later use, we calculate the maximized value of the likelihood function. Substituting (1.5.15) into (1.5.14), we obtain

$$\text{maximized log likelihood} = -\frac{n}{2} \log\left(\frac{2\pi}{n}\right) - \frac{n}{2} - \frac{n}{2} \log(SSR),$$

so that the maximized likelihood is

$$\max_{\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2} L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) = \left(\frac{2\pi}{n}\right)^{-n/2} \cdot \exp\left(-\frac{n}{2}\right) \cdot (SSR)^{-n/2}. \quad (1.5.16)$$

The Cramer-Rao Bound

Now, for the classical regression model (of Assumptions 1.1–1.5), the likelihood function $L(\tilde{\boldsymbol{\theta}})$ in the Cramer-Rao inequality is the conditional density (1.5.12), so the variance in the inequality is the variance conditional on \mathbf{X} . It can be shown that those regularity conditions are satisfied for the normal density (1.5.12) (see, e.g., Amemiya, 1985, Sections 1.3.2 and 1.3.3). We now calculate the information matrix for the classical regression model. The parameter vector $\boldsymbol{\theta}$ is $(\boldsymbol{\beta}', \sigma^2)'$. So $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\gamma})'$ and the matrix of second derivatives we seek to calculate is

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\theta}}'} = \begin{bmatrix} \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}'} & \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\gamma}} \\ \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\gamma} \partial \tilde{\boldsymbol{\beta}}'} & \frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\gamma}^2} \end{bmatrix}. \quad (1.5.17)$$

$\begin{matrix} (K \times K) & (K \times 1) \\ (1 \times K) & (1 \times 1) \end{matrix}$

The first and second derivatives of the log likelihood (1.5.13) with respect to $\tilde{\boldsymbol{\theta}}$, evaluated at the true parameter vector $\boldsymbol{\theta}$, are

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\beta}}} = \frac{1}{\gamma} \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (1.5.18a)$$

$$\frac{\partial \log L(\boldsymbol{\theta})}{\partial \tilde{\gamma}} = -\frac{n}{2\gamma} + \frac{1}{2\gamma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (1.5.18b)$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}'} = -\frac{1}{\gamma} \mathbf{X}'\mathbf{X}, \quad (1.5.19a)$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\gamma}^2} = \frac{n}{2\gamma^2} - \frac{1}{\gamma^3} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (1.5.19b)$$

$$\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\gamma}} = -\frac{1}{\gamma^2} \mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (1.5.19c)$$

Since the derivatives are evaluated at the true parameter value, we have $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\varepsilon}$ in these expressions. Substituting (1.5.19) into (1.5.17) and using $E(\boldsymbol{\varepsilon} | \mathbf{X}) = \mathbf{0}$ (Assumption 1.2), $E(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} | \mathbf{X}) = n\sigma^2$ (implication of Assumption 1.4), and recalling $\gamma = \sigma^2$, we can easily derive

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{1}{\sigma^2} \mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{n}{2\sigma^4} \end{bmatrix}. \quad (1.5.20)$$

Here, the expectation is conditional on \mathbf{X} because the likelihood function (1.5.12) is a conditional density conditional on \mathbf{X} . This block diagonal matrix can be inverted to obtain the Cramer-Rao bound:

$$\text{Cramer-Rao bound} \equiv \mathbf{I}(\boldsymbol{\theta})^{-1} = \begin{bmatrix} \sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0}' & \frac{2\sigma^4}{n} \end{bmatrix}. \quad (1.5.21)$$

Therefore, the unbiased estimator \mathbf{b} , whose variance is $\sigma^2 \cdot (\mathbf{X}'\mathbf{X})^{-1}$ by Proposition 1.1, attains the Cramer-Rao bound. We have thus proved

Proposition 1.6 (b is the Best Unbiased Estimator (BUE)): *Under Assumptions 1.1–1.5, the OLS estimator \mathbf{b} of $\boldsymbol{\beta}$ is BUE in that any other unbiased (but not necessarily linear) estimator has larger conditional variance in the matrix sense.*

This result should be distinguished from the Gauss-Markov Theorem that \mathbf{b} is minimum variance among those estimators that are unbiased *and* linear in \mathbf{y} . Proposition 1.6 says that \mathbf{b} is minimum variance in a larger class of estimators that includes nonlinear unbiased estimators. This stronger statement is obtained under the normality assumption (Assumption 1.5) which is not assumed in the Gauss-Markov Theorem. Put differently, the Gauss-Markov Theorem does not exclude the possibility of some nonlinear estimator beating OLS, but this possibility is ruled out by the normality assumption.

As was already seen, the ML estimator of σ^2 is biased, so the Cramer-Rao bound does not apply. But the OLS estimator s^2 of σ^2 is unbiased. Does it achieve the bound? We have shown in a review question to the previous section that

$$\text{Var}(s^2 | \mathbf{X}) = \frac{2\sigma^4}{n - K}$$

under the same set of assumptions as in Proposition 1.6. Therefore, s^2 does not attain the Cramer-Rao bound $2\sigma^4/n$. However, it can be shown that an unbiased estimator of σ^2 with variance lower than $2\sigma^4/(n - K)$ does not exist (see, e.g., Rao, 1973, p. 319).

The F -Test as a Likelihood Ratio Test

The **likelihood ratio test** of the null hypothesis compares L_U , the maximized likelihood without the imposition of the restriction specified in the null hypothesis, with L_R , the likelihood maximized subject to the restriction. If the likelihood ratio $\lambda \equiv L_U/L_R$ is too large, it should be a sign that the null is false. The F -test of the null hypothesis $H_0: \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ considered in the previous section is a likelihood ratio test because the F -ratio is a monotone transformation of the likelihood ratio λ . For the present model, L_U is given by (1.5.16) where the SSR , the sum of squared residuals minimized without the constraint H_0 , is the SSR_U in (1.4.11). The restricted likelihood L_R is given by replacing this SSR by the restricted sum of squared residuals, SSR_R . So

$$L_R = \max_{\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2 \text{ s.t. } H_0} L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) = \left(\frac{2\pi}{n}\right)^{-n/2} \cdot \exp\left(-\frac{n}{2}\right) \cdot (SSR_R)^{-n/2}, \quad (1.5.22)$$

and the likelihood ratio is

$$\lambda \equiv \frac{L_U}{L_R} = \left(\frac{SSR_U}{SSR_R}\right)^{-n/2}. \quad (1.5.23)$$

Comparing this with the formula (1.4.11) for the F -ratio, we see that the F -ratio is a monotone transformation of the likelihood ratio λ :

$$F = \frac{n - K}{\#\mathbf{r}} (\lambda^{2/n} - 1), \quad (1.5.24)$$

so that the two tests are the same.

Quasi-Maximum Likelihood

All these results assume the normality of the error term. Without normality, there is no guarantee that the ML estimator of β is OLS (Proposition 1.5) or that the OLS estimator \mathbf{b} achieves the Cramer-Rao bound (Proposition 1.6). However, Proposition 1.5 does imply that \mathbf{b} is a **quasi-** (or **pseudo-**) **maximum likelihood estimator**, an estimator that maximizes a misspecified likelihood function. The misspecified likelihood function we have considered is the normal likelihood. The results of Section 1.3 can then be interpreted as providing the finite-sample properties of the quasi-ML estimator when the error is incorrectly specified to be normal.

Conditional vs. Joint ML

Since a (joint) density is the product of a marginal density and a conditional density, the joint density of (\mathbf{y}, \mathbf{X}) can be written as

$$f(\mathbf{y}, \mathbf{X}; \zeta) = f(\mathbf{y} | \mathbf{X}; \theta) \cdot f(\mathbf{X}; \psi), \quad (1.5.25)$$

where θ is the subset of the parameter vector ζ that determines the conditional density function and ψ is the subset determining the marginal density function. For the linear regression model with normal errors, $\theta = (\beta', \sigma^2)'$ and $f(\mathbf{y} | \mathbf{X}; \theta)$ is given by (1.5.12).

Let $\tilde{\zeta} \equiv (\tilde{\theta}', \tilde{\psi}')'$ be a hypothetical value of $\zeta = (\theta', \psi)'$. Then the (full or joint) likelihood function is

$$f(\mathbf{y}, \mathbf{X}; \tilde{\zeta}) = f(\mathbf{y} | \mathbf{X}; \tilde{\theta}) \cdot f(\mathbf{X}; \tilde{\psi}). \quad (1.5.26)$$

If we knew the parametric form of $f(\mathbf{X}; \tilde{\psi})$, then we could maximize this joint likelihood function over the entire hypothetical parameter vector $\tilde{\zeta}$, and the ML estimate of θ would be the elements of the ML estimate of $\tilde{\zeta}$. We cannot do this for the classical regression model because the model does not specify $f(\mathbf{X}; \tilde{\psi})$. However, if there is no functional relationship between $\tilde{\theta}$ and $\tilde{\psi}$ (such as a subset of $\tilde{\psi}$ being a function of $\tilde{\theta}$), then maximizing (1.5.26) with respect to $\tilde{\zeta}$ is achieved by separately maximizing $f(\mathbf{y} | \mathbf{X}; \tilde{\theta})$ with respect to $\tilde{\theta}$ and maximizing $f(\mathbf{X}; \tilde{\psi})$ with respect to $\tilde{\psi}$. Thus, in this case of no functional relationship between $\tilde{\theta}$ and $\tilde{\psi}$, the conditional ML estimate of θ is numerically equal to the joint ML estimate of θ .

QUESTIONS FOR REVIEW

1. (Use of regularity conditions) Assuming that taking expectations (i.e., taking integrals) and differentiation can be interchanged, prove that the expected value of the score vector given in (1.5.5), if evaluated at the true parameter value θ , is zero. **Hint:** What needs to be shown is that

$$\int \frac{\partial \log f(\mathbf{y}; \theta)}{\partial \theta} f(\mathbf{y}; \theta) d\mathbf{y} = \mathbf{0}.$$

Since $f(\mathbf{y}; \tilde{\theta})$ is a density, $\int f(\mathbf{y}; \tilde{\theta}) d\mathbf{y} = 1$ for any $\tilde{\theta}$. Differentiate both sides with respect to $\tilde{\theta}$ and use the regularity conditions, which allows us to change the

order of integration and differentiation, to obtain $\int [\partial f(\mathbf{y}; \boldsymbol{\theta}) / \partial \tilde{\boldsymbol{\theta}}] d\mathbf{y} = \mathbf{0}$. Also, from basic calculus,

$$\frac{\partial \log f(\mathbf{y}; \boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}}} = \frac{1}{f(\mathbf{y}; \boldsymbol{\theta})} \frac{\partial f(\mathbf{y}; \boldsymbol{\theta})}{\partial \tilde{\boldsymbol{\theta}}}.$$

2. (Concentrated log likelihood with respect to $\tilde{\sigma}^2$) Writing $\tilde{\sigma}^2$ as $\tilde{\gamma}$, the log likelihood function for the classical regression model is

$$\log L(\tilde{\boldsymbol{\beta}}, \tilde{\gamma}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\tilde{\gamma}) - \frac{1}{2\tilde{\gamma}} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}).$$

In the two-step maximization procedure described in the text, we first maximized this function with respect to $\tilde{\boldsymbol{\beta}}$. Instead, first maximize with respect to $\tilde{\gamma}$ given $\tilde{\boldsymbol{\beta}}$. Show that the concentrated log likelihood (concentrated with respect to $\tilde{\gamma} \equiv \tilde{\sigma}^2$) is

$$-\frac{n}{2} [1 + \log(2\pi)] - \frac{n}{2} \log \left(\frac{(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})' (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})}{n} \right).$$

3. (Information matrix equality for classical regression model) Verify (1.5.10) for the linear regression model. **Hint:** If $\varepsilon_i \sim N(0, \sigma^2)$, then $E(\varepsilon_i^3) = 0$ and $E(\varepsilon_i^4) = 3\sigma^4$.
4. (Likelihood equations for classical regression model) We used the two-step procedure to derive the ML estimate for the classical regression model. An alternative way to find the ML estimator is to solve for the first-order conditions that set (1.5.18) equal to zero (the first-order conditions for the log likelihood is called the **likelihood equations**). Verify that the ML estimator given in Proposition 1.5 solves the likelihood equations.
5. (Maximizing joint log likelihood) Consider maximizing (the log of) the joint likelihood (1.5.26) for the classical regression model, where $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\beta}}', \tilde{\sigma}^2)'$ and $\log f(\mathbf{y} | \mathbf{X}; \tilde{\boldsymbol{\theta}})$ is given by (1.5.13). You would parameterize the marginal likelihood $f(\mathbf{X}; \tilde{\boldsymbol{\psi}})$ and take the log of (1.5.26) to obtain the objective function to be maximized over $\boldsymbol{\zeta} \equiv (\boldsymbol{\theta}', \boldsymbol{\psi}')'$. What is the ML estimator of $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \sigma^2)'$? [Answer: It should be the same as that in Proposition 1.5.] Derive the Cramer-Rao bound for $\boldsymbol{\beta}$. **Hint:** By the information matrix equality,

$$\mathbf{I}(\boldsymbol{\zeta}) = -E \left[\frac{\partial^2 \log L(\boldsymbol{\zeta})}{\partial \tilde{\boldsymbol{\zeta}} \partial \tilde{\boldsymbol{\zeta}}'} \right].$$

Also, $\partial^2 \log L(\boldsymbol{\zeta}) / (\partial \tilde{\boldsymbol{\theta}} \partial \tilde{\boldsymbol{\psi}}') = \mathbf{0}$.

References

Amemiya, T., 1985, *Advanced Econometrics*, Cambridge: Harvard University Press.

- Averch, H., and L. Johnson, 1962, "Behavior of the Firm under Regulatory Constraint," *American Economic Review*, 52, 1052–1069.
- Christensen, L., and W. Greene, 1976, "Economies of Scale in US Electric Power Generation," *Journal of Political Economy*, 84, 655–676.
- Christensen, L., D. Jorgenson, and L. Lau, 1973, "Transcendental Logarithmic Production Frontiers," *Review of Economics and Statistics*, 55, 28–45.
- Davidson, R., and J. MacKinnon, 1993, *Estimation and Inference in Econometrics*, Oxford: Oxford University Press.
- DeLong, B., and L. Summers, 1991, "Equipment Investment and Growth," *Quarterly Journal of Economics*, 99, 28–45.
- Engle, R., D. Hendry, and J.-F. Richards, 1983, "Exogeneity," *Econometrica*, 51, 277–304.
- Federal Power Commission, 1956, *Statistics of Electric Utilities in the United States, 1955, Class A and B Privately Owned Companies*, Washington, D.C.
- Jorgenson, D., 1963, "Capital Theory and Investment Behavior," *American Economic Review*, 53, 247–259.
- Koopmans, T., and W. Hood, 1953, "The Estimation of Simultaneous Linear Economic Relationships," in W. Hood, and T. Koopmans (eds.), *Studies in Econometric Method*, New Haven: Yale University Press.
- Krasker, W., E. Kuh, and R. Welsch, 1983, "Estimation for Dirty Data and Flawed Models," Chapter 11 in Z. Griliches, and M. Intriligator (eds.), *Handbook of Econometrics*, Volume 1, Amsterdam: North-Holland.
- Nerlove, M., 1963, "Returns to Scale in Electricity Supply," in C. Christ (ed.), *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford: Stanford University Press.
- Rao, C. R., 1973, *Linear Statistical Inference and Its Applications* (2d ed.), New York: Wiley.
- Scheffe, H., 1959, *The Analysis of Variance*, New York: Wiley.
- Wolak, F., 1994, "An Econometric Analysis of the Asymmetric Information, Regulator-Utility Interaction," *Annales D'Economie et de Statistique*, 34, 13–69.