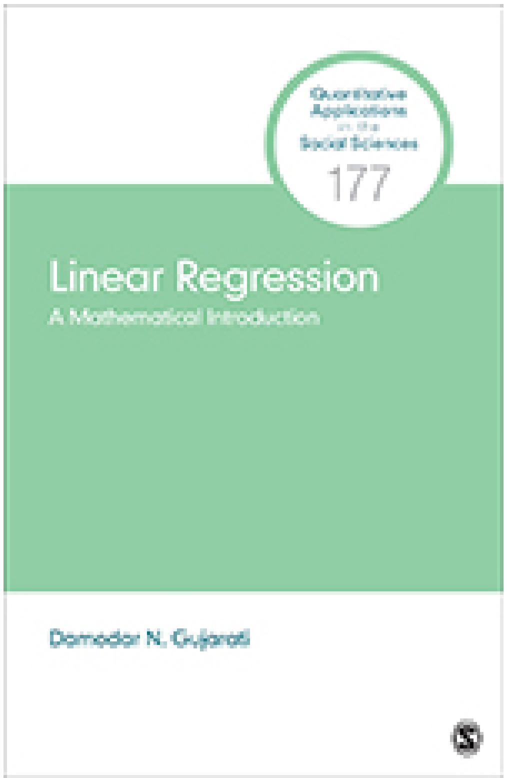

Quantitative Applications in the Social Sciences
177
Linear Regression
A Mathematical Introduction
Damodar N. Gujarati

Sage Research Methods

Linear Regression: A Mathematical Introduction

For the most optimal reading experience we recommend using our website.

https://methods.sagepub.com/book/mono/linear-regression/toc

Author: Damodar N. Gujarati

Pub. Date: 2019

Product: Sage Research Methods

DOI: https://doi.org/10.4135/9781071802571

Methods: Statistical inference, Linear regression, Regression analysis

Keywords: errors, parameters, equations, matrices, population, property

Disciplines: Sociology, Education, Psychology, Health, Social Policy and Public Policy, Political Science and

© 2019 by DNG Econometrics, Inc

International Relations, Geography, Criminology and Criminal Justice, Business and Management

Access Date: February 27, 2026

Publisher: SAGE Publications, Inc

City: Thousand Oaks

Online ISBN: 9781071802571

© 2019 SAGE Publications, Inc All Rights Reserved.

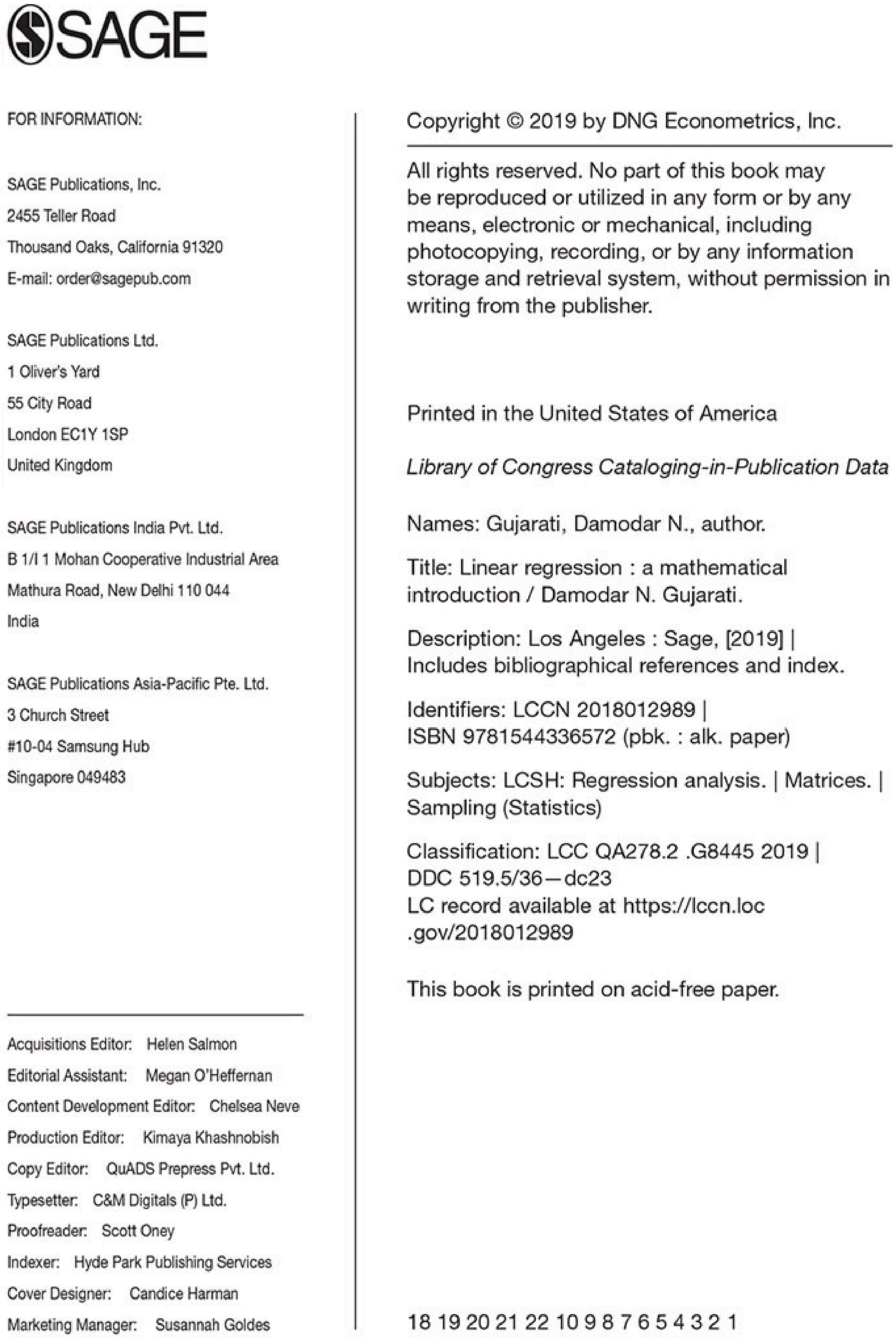
© 2019 by DNG Econometrics, Inc

Front Matter

Chapters

Back Matter

Copyright



SAGE

FOR INFORMATION:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks, California 91320
E-mail: order@sagepub.com

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Asia-Pacific Pte. Ltd.
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Acquisitions Editor: Helen Salmon
Editorial Assistant: Megan O'Heffernan
Content Development Editor: Chelsea Neve
Production Editor: Kimaya Khashnobish
Copy Editor: QuADS Prepress Pvt. Ltd.
Typesetter: C&M Digitals (P) Ltd.
Proofreader: Scott Oney
Indexer: Hyde Park Publishing Services
Cover Designer: Candice Harman
Marketing Manager: Susannah Goldes

Copyright © 2019 by DNG Econometrics, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Names: Gujarati, Damodar N., author.

Title: Linear regression : a mathematical introduction / Damodar N. Gujarati.

Description: Los Angeles : Sage, [2019] | Includes bibliographical references and index.

Identifiers: LCCN 2018012989 | ISBN 9781544336572 (pbk. : alk. paper)

Subjects: LCSH: Regression analysis. | Matrices. | Sampling (Statistics)

Classification: LCC QA278.2 .G8445 2019 | DDC 519.5/36—dc23
LC record available at https://lccn.loc .gov/2018012989

This book is printed on acid-free paper.

18 19 20 21 22 10 9 8 7 6 5 4 3 2 1

FOR INFORMATION:

SAGE Publications, Inc.

2455 Teller Road

Thousand Oaks, California 91320

E-mail: order@sagepub.com

SAGE Publications Ltd.

1 Oliver’s Yard

55 City Road

London EC1Y 1SP

United Kingdom

SAGE Publications India Pvt. Ltd.

B 1/I 1 Mohan Cooperative Industrial Area

Mathura Road, New Delhi 110 044

India

SAGE Publications Asia-Pacific Pte. Ltd.

3 Church Street

#10-04 Samsung Hub

Singapore 049483

Copyright © 2019 by DNG Econometrics, Inc.

All rights reserved. No part of this book may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission in writing from the publisher.

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

© 2019 by DNG Econometrics, Inc

Names: Gujarati, Damodar N., author.

Title: Linear regression : a mathematical introduction / Damodar N. Gujarati.

Description: Los Angeles : Sage, [2019] | Includes bibliographical references and index.

Identifiers: LCCN 2018012989 | ISBN 9781544336572 (pbk. : alk. paper)

Subjects: LCSH: Regression analysis. | Matrices. | Sampling (Statistics)

Classification: LCC QA278.2 .G8445 2019 | DDC 519.5/36—dc23 LC record available at https://lccn.loc.gov/2018012989

This book is printed on acid-free paper.

Acquisitions Editor: Helen Salmon

Editorial Assistant: Megan O’Heffernan

Content Development Editor: Chelsea Neve

Production Editor: Kimaya Khashnobish

Copy Editor: QuADS Prepress Pvt. Ltd.

Typesetter: C&amp;M Digitals (P) Ltd.

Proofreader: Scott Oney

Indexer: Hyde Park Publishing Services

Cover Designer: Candice Harman

Marketing Manager: Susannah Goldes

Dedication

Dedication

For Joan Gujarati, Diane Gujarati-Chesnut, Charles Chesnut, and my grandchildren, “Tommy” and Laura Chesnut, and for her behind-the-scenes help, Karen Low.

List of Figures

List of Figures

© 2019 by DNG Econometrics, Inc

Series Editor’s Introduction

Series Editor’s Introduction

I am very pleased to introduce *Linear Regression: A Mathematical Introduction* by Damodar Gujarati, one of the best-known econometricians of our era. The volume is a succinct introduction to the mathematics and statistical theory that is the foundation for classical linear regression analysis. It could be a course supplement for an advanced undergraduate or early graduate class in linear models. Alternatively, instructors might find it useful as a main text for the increasingly popular bootcamps in mathematics and statistics offered at the beginning of PhD programs. Even those already trained in these methods but in need of a refresher will find it of value. The volume is a welcome addition to the QASS Series.

Linear Regression: A Mathematical Introduction is very well structured and proceeds logically and carefully from the simplest to more complex linear regression models. When I read the draft manuscript, it was easy to imagine Professor Gujarati standing in front of the class working through the proofs and derivations on the chalkboard (or screen, as the case may be). As he would in person, Professor Gujarati explains how he proceeds from one step to the next, with lots of hints and tips. Reflecting its pedagogical purpose, there are exercises at the end of each chapter. Truth in advertising: the volume is mathematical, and readers who already know matrix algebra will get the most out of it. For readers who are a little rusty, there is an appendix that provides a brief but extremely helpful review. The appendix is also an entry point for readers without much training in matrix algebra. (A more comprehensive introduction can be found in *Matrix Algebra*, QASS Volume 38, by Krishnan Namboodiri.)

Importantly, *Linear Regression: A Mathematical Introduction* contains a clear exposition of the assumptions underlying the linear regression model, the consequences of violating each one, and the modifications of ordinary least squares needed to estimate linear regression models when this occurs. The volume provides a particularly helpful discussion of endogenous regressors, a perennial problem in social science applications. Indeed, the volume is remarkably practical. Once the key terms are defined and the foundations are laid, Professor Gujarati has plenty of practical advice about how, for example, to diagnose heteroscedasticity problems, interpret a Wald test, or assess the strength of an instrumental variable. With the advent of sophisticated software for statistical analysis, it is possible to run regression analyses without knowing the assumptions on which estimation and inference are based. Such ignorance is harmful. This volume is an antidote.

Classical linear regression analysis is one of the workhorses of the social sciences. Look at any major core journal in a social science discipline and you will find plenty of applications. At least as important is that linear regression is the foundation on which many more advanced statistical techniques are built.

Multilevel models, simultaneous equations, and structural equation models are just a few examples of techniques rooted in regression. A thorough understanding of classical regression is necessary for a thorough understanding of these and other statistical models. *Linear Regression: A Mathematical Introduction* helps build that foundation. Students of all ages and stages will benefit from this volume.

© 2019 by DNG Econometrics, Inc

Preface

Preface

Regression analysis is one of the most widely and intensively used techniques of quantitative research in fields as diverse as economics, finance, accounting, marketing, politics, international relations, agriculture, medicine, and biology. In fact, it is used in any area of research where one is interested in studying the relationship between a variable of interest, called the response variable, and a set of predictor variables. Sir Francis Galton (1822–1911) used it in the study of heredity, particularly the height of grown-up children in relation to the height of their parents. He used the method of least squares, the workhorse of linear regression analysis, for this purpose. Since then, the methodology of regression analysis has been improved and developed in many ways. It is no exaggeration to say that regression analysis has become an integral part of almost all scientific disciplines.

There is a well-developed mathematical and statistical theory behind the commonly used regression techniques. Some of this theory is quite complicated. My primary objective in writing this “Green Book” is to explain this theory in a rigorous but approachable manner to a large group of students, researchers, and teachers in various disciplines. With the ready availability of user-friendly statistical packages, estimating regression models is not a daunting task. But blindly using these packages without understanding the underlying theory could be a fruitless task, and at times, it could lead to misleading conclusions and policy prescriptions.

In about 250 pages, I explain the basics of linear regression, that is, regression linear in the parameters. The book contains seven chapters and four appendices. The key features of this book are as follows.

Key Features of the Book

1. A concise discussion of the ordinary least squares (OLS) and both the small- and large-sample properties of the OLS estimators.
2. A concise discussion of the method of maximum likelihood (ML) and the small- and large-sample properties of ML estimators.
3. A concise discussion of the distribution theory and a discussion of the commonly used tests of significance.
4. A concise discussion of the generalized least squares (GLS).
5. A concise discussion of the method of instrumental variables (IV) in cases where the

predictor variables are stochastic. The classical least-squares model assumes that the predictors are either independent or at least uncorrelated with the regression error term.

6. Four appendices on the basics of matrix algebra, essentials of large-sample theory, small- and large-sample properties of estimators, and important probability distributions.
7. Two extended examples that discuss the various methods discussed in the book.

The technical discussion of some of the topics is put in the appendices to the various chapters for the benefit of more advanced students.

For upper-level undergraduate students, this book will provide a solid introduction to the linear regression models. Graduate students, teachers, and researchers will find this book to be a quick reference for the major themes in linear regression analysis.

Two datasets to accompany the book are available on a website at: **study.sagepub.com/gujarati**.

© 2019 by DNG Econometrics, Inc

About the Author

About the Author

Damodar Gujarati (M.B.A. and Ph.D., both from University of Chicago) is Professor Emeritus of economics at the United States Military Academy at West Point. Prior to that, he taught for 25 years at the Baruch College of the City University of New York (CUNY) and at the Graduate Center of CUNY. He is the author of Government and Business, (McGraw Hill, 1984), the best-selling textbook Basic Econometrics (5th edition, 2009, with co-author Dawn Porter), as well as Essentials of Econometrics (4th edition, 2009, also with co-author Dawn Porter), both published by McGraw-Hill, and also Econometrics by Example (2nd edition, 2014, Palgrave-Macmillan). His experience spans business, consulting, and academia.

© 2019 by DNG Econometrics, Inc

Acknowledgments

Acknowledgments

I am grateful to the following reviewers for their incisive comments and suggestions on the draft of the book which has materially improved topics included in the book.

Michael Lewis-Beck, The University of Iowa
Colin Lewis-Beck, Iowa State University
Wendy Martinek, Binghamton University–SUNY
Kirk Randazzo, University of South Carolina
Jay Verkuilen, The Graduate Center–CUNY David Weakliem, University of Connecticut

During the preparation of the book I received valuable advice from the following academics:

Frank Fabozzi, EDHEC Business School, Paris Michael Grossman, The Graduate Center–CUNY
Charles B. Moss, University of Florida

I also want to thank Professors Paul M. Kellstedt and Guy D. Whitten at the Department of Political Science, Texas A&amp;M University, for their valuable guidance on the earlier drafts of the book.

I owe special thanks to Professor Randall Campbell of the College of Business at Mississippi State University for reading the manuscript and for checking the various mathematical and statistical formulas and correcting some of the errors.

For their behind-the-scenes help, I am thankful to Barbara Entwisle, Kenan Distinguished Professor of Sociology, University of North Carolina at Chapel Hill, and series editor of *Quantitative Applications in the Social Sciences* (QASS); to Helen Salmon, Senior Acquisitions Editor of Research Methods, Statistics and Evaluation at SAGE Publishing; to Megan OHeffernan, Editorial Assistant at SAGE Publishing; to Rajasree Ghosh at QuADS, for copyediting this book; and to Kimaya Khashnobish and Arnab Karmakar, Production Editors at SAGE publishing.

The Linear Regression Model (LRM)

1.1 Introduction

Regression analysis is one of the most widely and intensively used techniques of quantitative research in fields as diverse as economics, finance, accounting, marketing, politics, international relations, agriculture, medicine, and biology. In fact, it is used in any area of research where one is interested in studying the relationship between a variable of interest, called the response variable, and a set of predictor variables. Sir Francis Galton (1822–1911) used it in the study of heredity, particularly the height of grown-up children in relation to the height of parents. The regression technique that was used to study this relationship was the method of least squares. Since then the methodology of regression has been improved and developed in many ways. It is no exaggeration to say that regression analysis has become an integral part of research in almost all scientific disciplines.

Just to give a few examples, linear regression has been used in analyzing stock market returns, in the analysis of production and cost functions, in analyzing fertility and mortality rates, in the analysis of investment functions, in the analysis of the relationship between sugar intake and diabetes, in the analysis of death rates in relation to several factors, in the analysis of women's participation rates in the labor force, in the analysis of housing starts in relation to several socioeconomic variables, in the analysis of the effects of cigarette smoking on various types of cancer, in the analysis of admissions to graduate schools, in the analysis of presidential popularity, in the analysis of mental health, in the analysis of credit ratings of corporations, and in the analysis of crime rates in various suburban areas. In short, regression analysis has been used in a variety of situations, where the interest is in studying the relationship between the variable of interest in relation to several factors appropriate for the particular subject.

My primary focus in this book is on **linear regression**, which is the workhorse of regression analysis in most applications. Linear regression is also the foundation of the **generalized linear models (GLM)**, which I do not discuss in this book, for that requires a separate book.

In this and the following six chapters, I discuss the nature of linear regression and its theoretical foundations. Although students in applied disciplines may be interested in seeing how regression analysis is used in practice, they might benefit from learning about the underlying theory.

We express a generic linear regression model (LRM) as follows:

(1.1)
$$Y_i = B_1X_{1i} + B_2X_{2i} + B_3X_{3i} + \cdots + B_kX_{ki} + u_i \, i = 1, 2, 3, \ldots, n$$

In this model, Y is the dependent variable; alternative names are explained variable, predictand, **regressand**, response, endogenous variable, outcome, and controlled variable. In this book, we will use the term *regressand*, which is a rather neutral term.

In this model, X_1, X_2, . . . , X_k are called the explanatory variables. Alternative names are independent variable, predictor, **regressor**, stimulus, exogenous variable, covariate, and control variable. We will use the more neutral term *regressor*. Some of the regressors are quantitative, and some are qualitative, such as race, gender, religion, and nationality. Very often, such qualitative variables are represented by **dummy variables**, taking values of 1 or 0, with 1 indicating the presence of an attribute and 0 indicating its absence. Sometimes the dummy variables are multicategorical, as we will illustrate with a concrete example in Chapter 4.

The subscript *i* is the observation subscript. By convention, the subscript *i* is used if the data are **cross-sectional** and the subscript *t* is used if the data are **time series**. If the data involve both cross-section and time-series observations, we use the double subscript *it*, as in X_{kit}, meaning the *i*th and *t*th observations on the regressor X_k. The number of observations is denoted by *n.*

We call (1.1) an LRM, and the meaning of the term *linear* will be explained shortly.

Generally, the variable X_1 takes the value of 1 for each observation in the data. This is to allow for the intercept in the model. As a result, we can write (1.1) as

(1.2)
$$Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + \cdots + B_kX_{ki} + u_i$$

We call (1.2) a ***k*-variable regression model**. The actual value of *k* depends on the phenomenon of study. Initially, we assume that the values of the regressors are fixed. Given the fixed values of the regressors, we draw repeated samples of *Y* values. We call this setup the **fixed regressor case**. In Chapter 6, we will show what happens if the values of the *X* regressors are also drawn randomly. In this case, both the regressand and the regressors are drawn randomly. This is the case of the **stochastic regressor**. *Stochastic*, a Greek word, means relating to a process involving a randomly determined sequence of observations, each of which is considered as a sample of one element from a probability distribution.

Any model, however extensive, is not expected to explain the phenomenon of interest fully due to random or uncontrolled influences. To account for the unavoidable random variation, we add the random variable *u* to the model, which is called the **error term**. It represents all those factors that may affect the regressand but are not included in the model because their individual influence on the regressand is very small and collectively all these factors may cancel out each other. It is also called a **random error**. More accurately, it is called a **stochastic error term**. Note that the error term is also known as the **disturbance term**.

The coefficients, B_1, B_2, . . . , B_k are called the **regression parameters** or **coefficients**. In the LRM, it is assumed that the regression parameters are fixed numbers and not random, although we do not know their values.[1] Once we have a set of data, we will show how the values of the parameters are estimated. The coefficient B_1 is called the **intercept** and the coefficients B_2 through B_k are called the **partial regression coefficients**, for reasons to be discussed shortly. In practice, it is better to retain the intercept in the model, although there are situations where it may be suppressed, as will be discussed subsequently.

1.2 Meaning of “Linear” in Linear Regression

Before proceeding further, it is important to know the meaning of the term *linear*, for it can be interpreted in two different ways. The first and perhaps more “natural” meaning of linearity is that the dependent variable is a linear function of the explanatory variables as in (1.1) or (1.2). In this sense, the following regressions are not linear regressions:

$$Y_i = B_1 + B_2 X_1^2 + u_i \text{ or } Y_i = B_1 + B_2 \frac{1}{X_i} + u_i$$

The second interpretation of linearity is that the dependent variable is a linear function of the parameters, as in (1.1) and (1.2); here both these functions are linear in the variables as well as the parameters. But now consider the following functions:

$$Y_i = B_1 + B_2^2 X_i + u_i \text{ or } Y_i = \frac{1}{1 + e^{B_1 + B_2 X_i + u_i}}$$

These functions are not linear functions of the parameters; they are nonlinear functions of one or more parameters.

For our purpose, from now on the term *linear regression* will mean a regression that is linear in the parameters; it may or may not be linear in the explanatory variables. This does not mean the two preceding models cannot be estimated, but they require different estimation techniques, which are beyond the scope of this book.

Written out fully, Equation (1.2) represents the following set of equations:

(1.3)

$$\begin{aligned} Y_1 &= B_1 + B_2X_{21} + B_3X_{31} + \cdots + B_kX_{k1} + u_1 \\ Y_2 &= B_1 + B_2X_{22} + B_3X_{32} + \cdots + B_kX_{k2} + u_2 \\ &\vdots \\ Y_n &= B_1 + B_2X_{2n} + B_3X_{3n} + \cdots + B_kX_{kn} + u_n \end{aligned}$$

This system of equations can be written more compactly as

(1.4)

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{pmatrix} 1 & X_{21}X_{31} \cdots & X_{k1} \\ \vdots & \ddots & \vdots \\ 1 & X_{2n}X_{3n} \cdots & X_{kn} \end{pmatrix} \begin{bmatrix} B_1 \\ B_2 \\ B_3 \\ \vdots \\ B_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$n \times 1n \times kk \times 1n \times 1$

which we write as

(1.5)

$$y = XB + u$$

where

> ***y*** = *n* × 1 column vector of observations on the dependent variable
>
> ***X*** = *n* × *k* matrix of *n* observations on *k* regressors, which includes a regressor whose value is 1 for each observation. ***X*** is often called the **data matrix**.
>
> ***B*** = *k* × 1 column vector of the *k* unknown regression parameters
>
> ***u*** = *n* × 1 column vector of *n* errors or disturbances u_i

Note: We represent matrices by capital bold letters and vectors by lowercase bold letters.

Using the rules of matrix addition and multiplication, the reader should verify that the Equations (1.3) and (1.4) are equivalent. Equation (1.5) is the **matrix representation** of the LRM.[2]

Equation (1.5) is often called the **population regression model (PRM)**, for it purports to show the relationship between the regressand and the regressors in the population of interest of some phenomenon. The concept of population is general and refers to a well-defined entity (people, firms, cities, states, countries, etc.) that is the focus of a statistical or econometric analysis.

As Equation (1.5) shows, the PRM consists of two components: (1) a **deterministic component**, ***XB***, and (2) a **random component**, ***u***. As shown below, ***XB*** can be interpreted as the **conditional mean** of Y_i, $E(Y_i|\boldsymbol{X})$, conditional on the given ***X*** values. Given the values of the regressors, Equation (1.5) states that an individual *Y* value is equal to the mean value of the population of which it is a member, plus or minus a random term. Note that we are using ***X*** as shorthand for the ***X*** matrix, that is, values taken by the regressors included in the data matrix.

The regression coefficient B_1 gives the *mean or average* value of the regressand when the values of regressors X_2 through X_k are all set to zero for each observation. The regression coefficients, B_2 through B_k, as noted previously, are known as **partial regression coefficients**. Each partial regression coefficient gives *the rate of change in the mean value of the regressand* for a unit change in the value of the regressor associated with it, holding all other regressor values constant.

One rarely observes the whole population of interest. Usually, given the values of the regressors, we draw a random sample of *Y* values and estimate $E(Y_i|\boldsymbol{X})$. Based on this estimate, we try to infer the true value of $E(Y_i|\boldsymbol{X})$. This dual task of estimation and hypothesis testing is the essence of statistical inference in regression analysis and in the other areas of statistics.

1.3 Estimation of the LRM: An Algebraic Approach

If ***b*** is a *k* × 1 vector of estimators of ***B***, we can write the estimated model as

(1.6)

$$y = Xb + e$$

where ***e***, called the vector of **residuals**, is the sample counterpart of ***u***.

Based on a random sample of ***y*** and fixed ***X***, how do we estimate Equation (1.5)? That is, how do we estimate the parameters in ***B***? A commonly used method to estimate the regression parameters is the method of **ordinary least squares (OLS)**.[3] To explain this method, we rewrite the sample counterpart of Equation (1.2) as follows:

(1.7)

$$Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + \cdots + b_kX_{ki} + e_i$$

Equivalently, we can write Equation (1.7) as

(1.8)

$$e_i = Y_i - b_1 - b_2X_{21} - b_3X_{3i} - \cdots - b_kX_{ki}$$

Or, in matrix notation,

(1.9)

$$e = y - Xb$$

Equation (1.9) states that the residual is the difference between the actual *Y* values and the estimated *Y* values obtained from the regression model (1.2). Since we know the (sample) values of the regressand and the regressor, this error term depends on the values of the estimated parameters in the model. Therefore, our objective is to obtain values of the regression parameters that would make the sum of the residuals as small as possible, ideally zero. However, this is not a good criterion, for we can have some large positive residuals and some large negative residuals that could make the sum of residuals practically zero. Likewise, we could have small positive residuals and small negative residuals and the sum of residuals can also be zero.

To avoid the problem of the signs of the errors, instead of minimizing the sum of errors, in OLS we minimize the sum of squared residuals as follows:

(1.10)

$$\sum e_i^2 = \sum (Y_i - b_1 - b_2X_i - b_3X_{3i} - \cdots - b_kX_k)^2$$

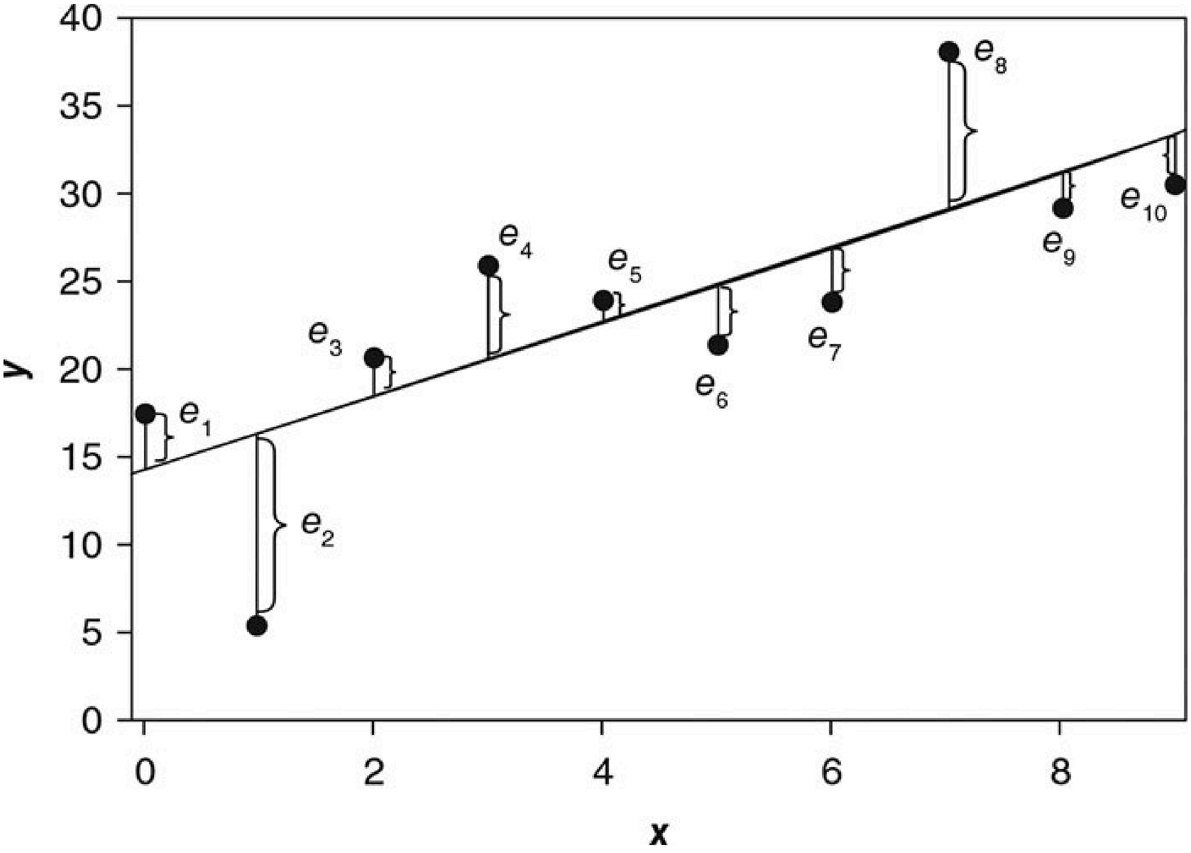
To get a visual picture of minimization of the squared residual sum of squares, consider Figure 1.1.

In Figure 1.1, we have hypothetical data on the dependent and explanatory variables, *Y* and *X*, respectively. Such a figure is known as a **scattergram**. The straight line shown in the figure is the regression line. Not all the data points shown in the figure lie on this regression line. The vertical distance between the data points

© 2019 by DNG Econometrics, Inc

and the regression line are the e_is, the residuals. The regression line is drawn in such a way that the sum of the squared residuals is as small as possible. How this is done is shown in what follows.

Figure 1.1 Hypothetical Scattergram



The minimization of the sum of squared residuals can be handled by calculus techniques. Specifically, we differentiate Equation (1.10) with respect to the unknown *b*s, set the resulting equations to zero (the first-order condition of optimization), and rearranging, we obtain *k* equations in *k* unknowns, known as the **normal equations** of least squares (see Appendix 1A).

In matrix notation, we can write Equation (1.10) as

(1.11)

$$\Sigma e_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$$

(1.12)

$$= \mathbf{y}'\mathbf{y} - 2\mathbf{bX}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{Xb}$$

© 2019 by DNG Econometrics, Inc

where use is made of the properties of the transpose of a matrix, namely, (***Xb***)′ = ***b*′*X*′**, and since ***b*′*X*′*y*** is a scalar (a real number), it is equal to its transpose ***y*′*Xb***. Note that $\sum e_i^2$ is called the **residual sum of squares (RSS)**.

To minimize ***e*′*e***, we first differentiate Equation (1.12) with respect to ***b*** to obtain

(1.13)

$$\frac{\partial \boldsymbol{e}'\boldsymbol{e}}{\partial \boldsymbol{b}} - 2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}$$

We now set these derivatives equal to zero (the first-order condition of optimization) to obtain the so-called **normal equations** of least squares:

(1.14)

$$X' - Xb = X'y$$

Since ***X*** is (*n* × *k*), ***X*′** is (*k* × *n*). Therefore, ***X*′*X*** is (*k* × *k*), a square matrix.

If the inverse of ***X*′*X*** exists, we can obtain

(1.15)

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

which reduces to

(1.16)

$$\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

This expression gives the estimators of the *k* unknown ***B*** coefficients. Notice that ***b*** is a **linear estimator**, that is, a linear function of the regressand ***y***, which is obvious from this equation. Since ***y*** is random and the *X*s are fixed, ***b*** is also random. ***B*** is, however, nonrandom. Note that Equation (1.16) gives the **point estimator** of each regression coefficient. That is, for a given sample, we obtain just one value, called the **estimate**, of each parameter included in the LRM. We can also obtain the **interval estimate** of a parameter, which gives a range that might include the true value of the parameter with certain probability. We will illustrate interval estimation in detail in Chapter 4.

Note: For the inverse of ***X*′*X*** to exist, the matrix ***X*** must be of full (column) rank, *k*. This requires that the number of observations, *n*, must be greater than the number of parameters estimated, *k* in our case. In this case, ***X*′*X*** is also of rank *k*.

To prove that ***b*** does minimize Equation (1.11), we can differentiate Equation (1.16) with respect to ***b***, which yields

(1.17)

$$\frac{\partial \boldsymbol{e}'\boldsymbol{e}}{\partial \boldsymbol{b} \partial \boldsymbol{b}'} = 2\boldsymbol{X}'\boldsymbol{X}$$

If the matrix ***X*** has full rank (i.e., rank *k*), the matrix of second-order (partial) derivatives given in Equation (1.17), called the **Hessian matrix**, is a **positive definite (PD) matrix**,[4] thus establishing that the ***b*** given in Equation (1.16) is indeed a minimum of Equation (1.12).

Note that (***X'X***) is a symmetric square matrix of order *k* × *k.* Written explicitly, it has the following form:

(1.18)

$$(\boldsymbol{X}'\boldsymbol{X}) = \begin{pmatrix} n & \sum X_{2i} & \sum X_{3i} & \dots & \sum X_{ki} \\ \sum X_{2i} & \sum X_{2i}^2 & \sum X_{2i}X_{3i} & \dots & \sum X_{2i}X_{ki} \\ \sum X_{3i} & \sum X_{3i}X_{2i} & \sum X_{3i}^2 & \dots & \sum X_{3i}X_{ki} \\ \vdots & & & & \\ \sum X_{ki} & \sum X_{ki}X_{2i} & \sum X_{ki}X_{3i} & \dots & \sum X_{ki}^2 \end{pmatrix}$$

Notice these features of the (***X'X***) matrix: (1) it is a symmetric matrix—the entries on either side of the main diagonal (running from upper left to lower right) are mirror images of one another; (2) the first row and the first column of this matrix give the sums of regressor values; note that *n*, the sample size, is the number of 1s added; and (3) the entries on the main diagonal give the sums of squares of the ***X*** variables and the off-diagonal entries give the sums of pairwise products of the regressors. As noted, the matrix (***X'X***) is of the same rank as the matrix ***X***.

We now express the estimated regression function as

(1.19)

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb}$$

which says ***y***-hat or ***y***-caret is the estimated (population) mean value of the dependent variable, given the values of the regressors. In other words, ***Xb*** is an estimator of ***XB***.

© 2019 by DNG Econometrics, Inc

Recall that

(1.20)

$$\begin{aligned} e &= y - Xb \\ &= y - X(X'X)^{-1}X'y, \text{ using Equation(1.16)} \\ &= My \end{aligned}$$

where

(1.21)

$$M = I - X(X'X)^{-1}X'$$

where ***I*** is an (*n* × *n*) identity matrix, that is, a matrix with 1s on the main diagonal and zero elsewhere.

The ***M*** matrix has the following properties: It is square, symmetric ($\boldsymbol{M'} = \boldsymbol{M}$), idempotent (i.e., $\boldsymbol{M}^2 = \boldsymbol{M}$), singular, and of order *n* and rank (*n* − *k*), and it has the property that $\boldsymbol{MX} = \mathbf{0}$ (verify this).[5] As a result, we have

(1.22)

$$X'e = X'My = \mathbf{0}$$

which shows that the regressors and the residuals are orthogonal. In words, each regressor and the associated residuals are independent. Note that ***M*** is *singular*, as its rank is (*n* − *k*)—the number of observations minus the number of regressors. The proof of this will be presented shortly.

The ***M*** matrix is also known as a **projection matrix**.

We can now write Equation (1.19) as

(1.23)

$$\hat{\boldsymbol{y}} = \boldsymbol{Xb} = \boldsymbol{Hy}$$

where

(1.24)

$$H = X(X'X)^{-1}X'$$

H is called the “hat matrix” as it transforms ***y*** into ***y***-hat (ŷ). Like ***M***, ***H*** is also a **projection matrix**. Geometrically, it projects ***y*** (perpendicularly) onto ŷ (***y***-hat).

It is easy to verify that

(1.25)
$$H' = H,\ H^2 = H,\ H + M = I,\ \text{and } HM = \mathbf{0}$$

As a result,

(1.26)
$$y = \boldsymbol{Hy} + \boldsymbol{My} = \hat{\boldsymbol{y}} + \boldsymbol{e} = \boldsymbol{Xb} + \boldsymbol{e}$$

Furthermore, because of Equations (1.20) and (1.22), it follows that

(1.27)
$$\hat{y}\boldsymbol{e} = 0$$

That is, the vectors *y*-hat and ***e*** are orthogonal. In other words, the estimated *Y* values and the residuals are independent.

We can visualize this relation in Figure 1.2.

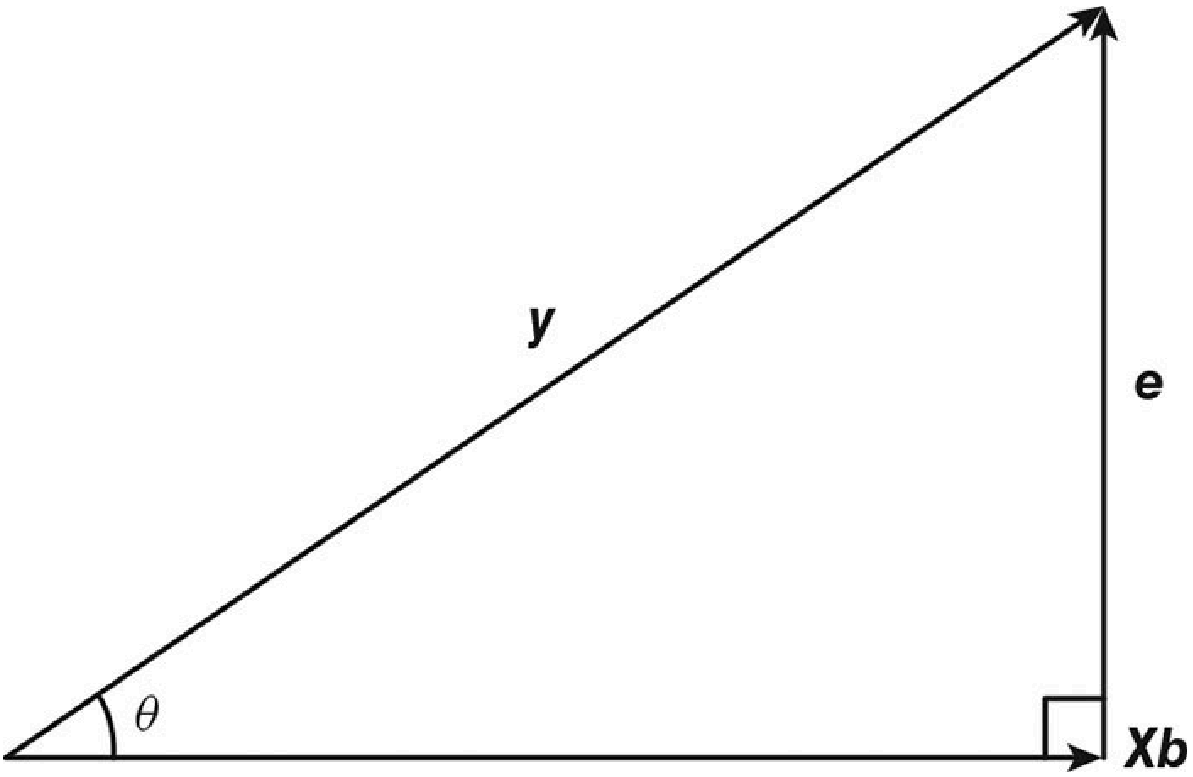
Figure 1.2 shows the vector ***y***, the estimated ***y*** vector (= ***Xb***), and the residual vector, ***e***. As you can see, the residual vector (***e***) makes a right angle with the estimated ***y*** vector. This figure is an example of an **orthogonal projection**, or loosely speaking "dropping a perpendicular." Technically speaking, "a projection is a mapping that takes each point of E^n into a point in a subspace of E^n, while leaving all points in that subspace unchanged. Because of this, the subspace is called the **invariant subspace** of the projection."[6] The projection matrices discussed above perform this orthogonal projection.

In short, the matrix ***H*** applied to ***y*** gives the vector of fitted or estimated values of ***y*** and the matrix ***M*** applied to ***y*** gives the vector of least-squares residuals, ***e***.

Note an interesting relationship between the projection matrices ***H*** and ***M***:

(1.28)
$$H + M = I$$

Figure 1.2 Orthogonal Projection (Residuals and Fitted Values)



These two projection matrices are therefore **complementary projections** because

(1.29)

$$Hy + My = y$$

That is, ***H*** and ***M*** applied to ***y*** produce the original vector ***y***, as can be readily verified.

Another interesting property of the two projection matrices is that their product produces the **null matrix 0**:

(1.30)

$$HM = \mathbf{0}$$

In other words, the two projection matrices **annihilate** each other.[7] Every element of the null matrix is zero.

The matrices ***M*** and ***H*** play a vital role in analyzing LRMs, as the subsequent discussion will show.

© 2019 by DNG Econometrics, Inc

1.4 Goodness of Fit of a Regression Model: The Coefficient of Determination (R^2)

Besides estimating the parameters of a regression model, we are often interested in finding how well the chosen regression model explains the variation in the regressand. For this purpose, we obtain a measure of goodness of fit, called the **coefficient of determination**, denoted by R^2. It measures the proportion or the percentage of variation in ***y*** explained by the regression model—that is, by the regressors included in the model.

To explain R^2, we develop the concepts of **total sum of squares** (**TSS**), **explained sum of squares** (**ESS**), and the RSS. Once these quantities are estimated, we define R^2 as

(1.31)

$$R^2 = \frac{\text{Explained sum of squares}}{\text{Total sum of squares}} = \frac{\text{ESS}}{\text{TSS}}$$

We start with Equation (1.26):

(1.26)

$$y = Hy + My = \hat{y} + e = Xb + e$$

Premultiplying both sides of this equation by ***y′***, we obtain

(1.32)

$$\begin{aligned} y'y &= (Xb+e)'(Xb+e) \\ &= b'X'Xb + e'(Xb+e) \\ &= b'X'Xb + e'e, \ \text{since } X'e = 0 \end{aligned}$$

However, this is the raw sum of squares of the actual Y values, $\boldsymbol{y}'\boldsymbol{y} = \Sigma Y_i^2$.

Verbally, we can write Equation (1.32) as

raw TSS = raw ESS + raw RSS

where TSS = total sum of squares; ESS = explained sum of squares, that is, that part of TSS explained by

© 2019 by DNG Econometrics, Inc

the regression model; and RSS = residual sum of squares, that is, that part of TSS not explained by the regression model.

The term *raw* means these sums of squares are not measured as deviations from their respective mean values.

If the regression model contains an intercept term, we would like to compute the sum of squares of *Y* values around its mean value, that is,

(1.33)

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma Y_i^2 - n\bar{\boldsymbol{Y}}^2 = \boldsymbol{y}'\boldsymbol{y} - n\bar{\boldsymbol{Y}}^2$$

Subtracting,$\boldsymbol{n}\bar{\boldsymbol{Y}}^2$ from both sides of Equation (1.32), we obtain

(1.34)

$$\begin{aligned} (\boldsymbol{y}'\boldsymbol{y} - n\bar{\boldsymbol{Y}}^2) &= (\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} - n\bar{\boldsymbol{Y}}^2) + \boldsymbol{e}'\boldsymbol{e} \\ \text{TSS} &= \text{ESS} + \text{RSS} \end{aligned}$$

TSS is the mean-corrected total sum of squares of the regressand; ESS is the mean-corrected explained sum of squares, that is, that part of the TSS that is explained by the regressors in the model; and RSS, the residual sum of squares, is the remainder of TSS that is not explained by the regressors. If the estimated model fits the data well, we would expect ESS to explain a substantial variation in the regressand. The coefficient of determination is a measure of how well the fitted model explains the variation in the dependent variable.

Following Equation (1.31), we obtain

(1.35)

$$R^2 = \frac{\boldsymbol{b}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} - n\bar{\boldsymbol{Y}}^2}{\boldsymbol{y}'\boldsymbol{y} - n\bar{\boldsymbol{Y}}^2}$$

R^2 is also defined as

(1.36)

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'\boldsymbol{y} - n\bar{\boldsymbol{Y}}^2}$$

The R^2 thus defined lies between 0 and 1.

It is a property of R^2 that it increases as additional regressors are added to the model. To compare regression models with the *same* regressand but differing number of regressors, researchers use a variant of R^2 known as the **adjustedR^2**, denoted by $\bar{R}^2$ (read as R-bar squared).

It is defined as follows:

(1.37)

$$\bar{R}^2 = 1 - \frac{\Sigma e_i^2 \big/ (n-k)}{\Sigma(Y_i - \bar{Y}^2) \big/ (n-1)}$$

This can be translated as follows:

(1.38)

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

which not only shows the relationship between adjusted and unadjusted R^2 but also shows that for k &gt; 1, $\bar{R}^2$ &lt; R^2, which implies that as the number of regressors in the model increases, the adjusted R^2 increases less than the unadjusted R^2. The term *adjusted* means adjusted for the **degrees of freedom** (*df*)[8] for the sums of squares associated in Equation (1.34). So to speak, there is a penalty for adding more regressors to the model. Adjusted R^2 is often used to compare competing regression models. But it is important to note that to compare two or more regression models on the basis of $\bar{R}^2$, *the dependent variable must be the same.* Thus, if the dependent variable in one model is Y but log Y in another model, we cannot use $\bar{R}^2$ to compare the two models. This is because variation in Y and the variation in log Y are not the same. In the former, it is the absolute change, whereas in the latter it is the relative change.

A caution on the role of R^2 in regression analysis: The late Professor Arthur Goldberger has this comment about the R^2.

> From our perspective, R^2 has a very modest role in regression analysis, being a measure of goodness of fit of a sample LS [least square] linear regression in a body of data. Nothing in the CR [classical regression] model requires that R^2 be high. Hence a high R^2 is not evidence in favor of the model, and a low R^2 is not evidence against it. . . . In fact the most important thing about R^2 is that

> it is not important in the CR model. The CR model is concerned with parameters in a population, not with goodness of fit in the sample.[9]

However, R^2 and adjusted R^2 have become a standard feature of most statistical packages.

We discuss the classical LRM in Chapter 2.

1.5 R^2 for Regression Through the Origin

Occasionally, we come across a regression model without the constant, or intercept, term. Such a model is called **regression through the origin** or **zero-intercept model**. For this model, we use the raw sums of squares defined in Equation (1.32) and obtain what is called the **raw**R^2.

(1.39)

$$R^2_{\text{raw}} = \frac{\text{ESS}_{\text{raw}}}{\text{TSS}_{\text{raw}}}$$

The raw R^2 does not have the same properties as the traditional R^2. But we will have more to say about this measure in Chapter 2. *Suffice it to note here that unless there is a compelling reason to use the zero-intercept model, it is better to retain the intercept in practice.*

1.6 An Example: The Determination of Hourly Wages in the United States

Before we conclude this chapter, here is a concrete example of a linear regression. Based on a random sample of 1,289 workers from the Current Population Survey (CPS) for March 1995, we obtained the following regression based on the method of OLS.

(1.40)

$$\begin{aligned} W_i &= -7.1833 - 3.0748FE_i - 1.5653NW_i + 1.0959UN_i + 1.3703ED_i + 0.1666EX_i \\ n &= 1,289, \quad R^2 = 0.3233 \end{aligned}$$

where *W* = hourly wage in dollars, *FE* (gender), coded 1 for female, 0 for male, *NW* (race), coded 1 for non-white workers and 0 for white workers, *UN* (union status), coded 1 if in union job, 0 otherwise, *ED* (education) in years, *EX* (work experience) in years. In this regression, *FE*, *NW*, and *UN* are **dummy variables** and *ED* and *EX* are quantitative variables. For discussion purposes, we call (1.40) a **wage regression**. Note that *EX*, the experience variable, is defined as age minus years of schooling minus 6; it is assumed that schooling starts at 6 years of age.

The details of regression (1.40) will be fully discussed in Chapter 4. But this is how we interpret this regression. The negative intercept value in this regression has no viable economic meaning, for literally interpreted, it suggests that if the values of all the regressors in this regression are held constant at zero, the average wage is a negative $7.18. In many regressions, the intercept value may not be meaningful.

The interpretation of the quantitative variables is straightforward. For example, the education coefficient suggests that if education increases by a year, the average wage goes up by about $1.37, ceteris paribus (holding the values of the other regressors in the model constant). Similarly, the coefficient of the experience variable suggests that the average wage goes up by about $0.16 per year of service experience, ceteris paribus. The positive value of these two coefficients makes economic sense.

The coefficients of the qualitative, or dummy, variables are interpreted differently. Here the comparison is between the variable that gets a value of 1 and the one that gets the value of 0. Thus, the negative coefficient of the gender variable suggests that female workers, on average, earn less than male workers by about $3. Values of other dummy coefficients should be interpreted similarly. Again, note that the negative coefficients of gender and race variables and the positive coefficient of the union variable are in accordance with labor market behavior.

The R^2 value of about 0.32 suggests that the predictors or regressors included in the wage regression explain about 32% of the variation in the average wage. This value might seem low, but we will have more to say about this in Chapter 4, and also note the caution sounded by Arthur Goldberger.

1.7 Summary

In this chapter, we introduced the *k*-variable LRM in its most general form. At the outset, we stated that by an LRM we mean a regression model that is linear in the parameters and not necessarily linear in the variables.

We also discussed briefly the nature of the *stochastic* error term u_i and stated that it is an integral part of the LRM. We introduced the LRM both in the scalar form and in the matrix form. Once we go beyond the simple two- or three-variable regression models, we need to use matrix algebra to avoid lengthy algebraic equations and derivations.

In practice, we rarely observe the true population of interest. Invariably, we believe we have a random sample from the population of interest, and all the analysis is based on the sample data. For this purpose, we introduced the **sample regression model**, which we use to estimate the unknown population parameters. To take into account sampling variability, we introduced the sample error term, called the residual, e_i, as a proxy for the true error term, u_i.

To estimate the unknown population parameters, we used the method of OLS. In OLS, we minimize the RSS, Σe_i^2. The resulting estimators of the regression parameters are known as **OLS estimators**.

Two of the important properties of OLS are that the estimated Y values and the residuals are uncorrelated and that each regressor is uncorrelated with the respective residual, assuming the $\boldsymbol{X}$ are fixed or constant. As we will show in the next chapter, the mean value of the residuals $\bar{e}$ is zero.

It is interesting that so far we have not made any assumptions regarding the probabilistic properties of the error term, u_i. The only assumptions we have made are that the $\boldsymbol{X}$ matrix is nonstochastic and that the number of observations, n, is greater than the number of parameters estimated, which is another way of saying that the data matrix $\boldsymbol{X}$ is of full (column) rank.

The purely algebraic approach to estimate regression parameters is not adequate, for the objective of regression analysis is not only to estimate the parameters of the PRM, which is based on sample data, but also to draw inferences about the true values of the parameters. Estimates based on sample data are subject to variation from sample to sample.

To accomplish the twin objectives of estimation and inference, we need a framework. The **classical linear regression model (CLRM)** provides such a framework. We discuss the CLRM in Chapter 2.

Exercises

1.1 Which of the following models are linear in the parameters, or variables, or both. Which of these models are LRMs?

a. $Y_i = B_1 + B_2 (1/X_i) + u_i$ (Reciprocal)

b. $Y_i = B_1 + B_2 \ln X_i + u_i$ (Semilogarithmic)

c. $\ln Y_i = B_1 + B_2 X_i + u_i$ (Inverse semilogarithmic)

d. $\ln Y_i = B_1 + B_2 \ln X_i + u_i$ (Double logarithmic)

e. $\ln Y_i = B_1 + B_2 (1/X_i) + u_i$ (Logarithmic reciprocal)

Note: = natural log and u_i is the regression error term.

1.2 Are the following models LRMs? Why or why not?

a. $Y_i = e^{B_1 + B_2 X_i + u_i}$

b. $Y_i = \frac{1}{1 + e^{B_1 + B_2 X_i + u_i}}$

c. $\ln Y_i = B_1 + B_2 (1 / X_i) + u_i$

d. $Y_i = B_1 + (0.75 - B_1) e^{-B_2 (X_i - 2)} + u_i$

e. $Y_i = B_1 + B_2^3 X_i + u_i$

1.3 Consider the following regression model that has no explanatory variables.

$Y_i = B_1 + u_i$

a. Use OLS to estimate B_1.

b. How would you interpret B_1 in this model?

1.4 Consider the following simple two-variable, or bivariate, regression model.

$Y_i = B_1 + B_2 X_i + u_i$

a. Using OLS, obtain the estimators of B_1 and B_2.

b. How would you interpret the two regression coefficients?

1.5 Prove: $r^2_{Y\hat{Y}} = R^2$, that is, the squared correlation coefficient between the actual Y and the estimated Y from a regression model is equal to the coefficient of determination.

1.6 Let B_{yx} be the slope coefficient in the regression of Y on X and B_{XY} the slope coefficient in the regression of X on Y. Show that $B_{YX} B_{XY} = r^2$.

1.7 Show that $\bar{\boldsymbol{X}} = \boldsymbol{1}'\boldsymbol{x} / n$. Similarly, show that $\bar{\boldsymbol{Y}} = \boldsymbol{1}'\boldsymbol{y} / n$. Show that $\sum_1^n \boldsymbol{X}_i = \boldsymbol{1}'\boldsymbol{x}$, where **1′** = (1, 1, . . . , 1)′ and $\boldsymbol{x}$ is an n-element column vector.

1.8 Prove the following in inequality, known as the **Cauchy–Schwarz inequality**:

$[E(XY)]^2 \le E(X^2) E(Y^2)$

Use this inequality to show that r^2, the squared correlation coefficient, is such that $0 \leq r^2 \leq 1$.

1.9 Show that

a. $\bar{e} = 0$, that is, the average value of the residuals is zero.

b. $\bar{Y} = b_1 + b_2\bar{X}_2 + b_3\bar{X}_3 + \cdots + b_k\bar{X}_k$, that is, the regression hyperplane passes through the sample mean values of Y and the Xs.

c. The mean value of actual Y and estimated Y values are the same.

1.10 Consider the following regression model:

$Y_i^k = B_1 + B_2X_i + u_i$ Y &gt; 0, $k = $ a constant

Consider the following values for k

$k = 1, 2, 0.5, -0.5, -1$

a. For each of these k values, find the corresponding regression model. Which of these models are LRMs?

b. Suppose $k = 0$. Can you estimate the regression model in this case?[10]

1.11 You are given 10 values for variables ***Y*** (the dependent variable) and ***X*** (the explanatory variable):

Y	70	65	90	95	110	115	120	140	155	150
X	80	100	120	140	160	180	200	220	240	260

Based on these values, estimate the following regression:

$Y_i = B_1 + B_2X_i + u_i$ $i = 1, 2, \ldots, 10$

a. Find $(\boldsymbol{X'X}),(\boldsymbol{X'Y})$

b. Estimate $\boldsymbol{b}=(\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}.$

c. Estimate R^2 for this example.

Appendix 1A: Derivation of the Normal Equations

We start with the following equation:

(1A.1)

$$\Sigma u_i^2 = \Sigma \left(Y_i - B_1 - B_2X_{2i} - B_3X_{3i} - \cdots - B_kX_{ki}\right)^2$$

Differentiating this equation partially with respect to each of the *B* coefficients, and setting them equal to 0, we obtain

(1A.2)

$$\begin{aligned}
\frac{\partial \Sigma u_i^2}{\partial B_1} &= 2\Sigma(Y_i - B_1 - B_2X_{2i} - B_3X_{3i} - \cdots - B_kX_{ki})(-1) = 0\\
\frac{\partial \Sigma u_i^2}{\partial B_2} &= 2\Sigma(Y_i - B_1 - B_2X_{2i} - B_3X_{3i} - \cdots - B_kX_{ki})(-X_{2i}) = 0\\
\frac{\partial \Sigma u_i^2}{\partial B_3} &= 2\Sigma(Y_i - B_1 - B_2X_{2i} - B_3X_{3i} - \cdots - B_kX_{ki})(-X_{3i}) = 0\\
&\vdots\\
\frac{\partial \Sigma u_i^2}{\partial B_k} &= 2\Sigma(Y_i - B_1 - B_2X_{2i} - B_3X_{3i} - \cdots - B_kX_{ki})(-X_{ki}) = 0
\end{aligned}$$

We have *k* equations in *k* unknown ***B*** coefficients and *in general* we can obtain the values of the *k* unknown regression parameters.[11] Replacing the unknown, and unobservable, ***B*** by the observable (or computed) ***b***, and simplifying, we obtain

(1A.3)

$$\begin{aligned}
nb_1 + b_2\Sigma X_{2i} + b_3\Sigma X_{3i} + \cdots + b_k\Sigma X_{ki} &= \Sigma Y_i\\
b_1\Sigma X_{2i} + b_2\Sigma X_{2i}^2 + b_3\Sigma X_{2i}X_{3i} + \cdots + b_k\Sigma X_{2i}X_{ki} &= \Sigma X_{2i}Y_i\\
b_1\Sigma X_{3i} + b_2\Sigma X_{3i}X_{2i} + b_3\Sigma X_{3i}^2 + \cdots + b_k\Sigma X_{3i}X_{ki} &= \Sigma X_{3i}Y_i\\
b_1\Sigma X_{ki} + b_2\Sigma X_{ki}X_{2i} + b_3\Sigma X_{ki}X_{3i} + \cdots + b_k\Sigma X_{ki}^2 &= \Sigma X_{ki}Y_i
\end{aligned}$$

Notice that each estimated ***B*** coefficient is expressed in terms of the observable squares and cross-products of the *X* and *Y* variables.

In matrix notation, Equation (3) can be expressed as

(1A.4)

$$\begin{pmatrix} n\ \Sigma X_{2i}\ \Sigma X_{3i}\ \cdots\ \cdots\ \cdots\ \cdots\ \Sigma X_{ki} \\ \Sigma X_{2i}\ \Sigma X_{2i}^2\ \Sigma X_{2i}X_{3i}\ \cdots\ \cdots\ \Sigma X_{2i}X_{ki} \\ \Sigma X_{3i}\ \Sigma X_{3i}X_{2i}\ \Sigma X_{3i}^2 \cdots\ \cdots\ \cdots\ \Sigma X_{3i}X_{ki} \\ \\ \Sigma X_{ki}\ \Sigma X_{ki}X_{2i}\ \Sigma X_{ki}X_{3i} \cdots\ \cdots\ \Sigma X_{ki}^2 \end{pmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \\ b_k \end{bmatrix} = \begin{bmatrix} 1\ 1\ 1\ 1 \ldots\ldots\ldots\ldots 1 \\ X_{21}X_{22}X_{23}\ldots\ldots\ldots X_{2n} \\ X_{31}\ X_{32}\ X_{33}\ldots\ldots\ldots X_{3n} \\ \\ X_{k1}\ X_{k2}\ X_{k3}\ldots\ldots\ldots X_{kn} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \\ Y_n \end{bmatrix}$$

$$(X'X)b \quad X'y$$
$$(k \times k)(k \times 1)(k \times n)(n \times 1)$$

Verify that the matrices and vectors follow the rules of matrix algebra. More compactly, Equation (4) can be expressed as

(1A.5)
$$(X'X)b = X'y$$

Since $\boldsymbol{X}$ is of rank k, and the inverse of ($\boldsymbol{X'X}$) exists, denoted by $(\boldsymbol{X'X})^{-1}$, we can premultiply Equation (5) on both sides by this inverse to obtain

(1A.6)
$$(X'X)^{-1}(X'X)b = (X'X)^{-1}(X'y)$$

Since $(X'X)^{-1}(X'X) = I$, an identity matrix of order ($n \times n$), we obtain

(1A.7)
$$b = (X'X)^{-1}(X^{'}y)$$
$$(k \times 1)\ (k \times k)(k \times n)\ (n \times 1)$$

This is a fundamental result of OLS regression.

It may be noted that $\boldsymbol{b}$ given in Equation (1A.7) is an example of a **linear estimator**.

It is of the form $\boldsymbol{Ly}$, where $\boldsymbol{L}$ is a matrix of real numbers; in the present case,

(1A.8)
$$L = (X'X)^{-1}X'$$

© 2019 by DNG Econometrics, Inc

Notes

[1] Followers of *Bayesian statistics* regard these parameters as random. In this book, we will not pursue Bayesian regression models. See, for example, Koop, G. (2003). *Bayesian econometrics*. Chichester, England: Wiley; Weakliem, D. L. (2016). *Hypothesis testing and model selection in the social sciences* (Chapter 4). New York, NY: Guilford Press.

[2] For the benefit of readers who are not familiar with matrix algebra or whose knowledge of the subject has become a little rusty, Appendix A provides a summary of the major themes in matrix algebra.

[3] OLS is a special case of the method of **generalized least squares**, which we will discuss in Chapter 5.

[4] A symmetric matrix $\boldsymbol{A}$ is a positive definite matrix if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x}$ &gt; 0 for all nonzero $\boldsymbol{x}$. It is a positive semidefinite matrix if $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} \geq 0$ for all $\boldsymbol{x}$ and there is at least one nonzero $\boldsymbol{x}$ for which $\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x} = 0$. See Appendix A for further details.

[5] For any idempotent matrix, its rank is equal to its trace. Here, $\mathrm{tr}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \mathrm{tr}(\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) = \mathrm{tr}(\boldsymbol{I}_k) = k$. Hence $\mathrm{tr}(\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{I}_n) - k = n - k$. That is why $\boldsymbol{M}$ is singular. Remember that a matrix whose determinant value is 0 is called a singular matrix.

[6] Davidson, R., &amp; MacKinnon, J. (2004). *Econometric theory and methods* (p. 57). New York, NY: Oxford University Press.

[7] Davidson, R., &amp; MacKinnon, J. (2004). *Econometric theory and methods* (p. 59). New York, NY: Oxford University Press.

[8] The degrees of freedom means the number of values free to vary when computing a statistic. For example, the mean value of five observations [5, 8, 10, 13, 14] is 10. To keep the mean value at 10, we can change the value(s) of only four number(s). So here we have only 4 degrees of freedom, although there are 5 observations.

[9] Goldberger, A. S. (1991). *A course in econometrics* (p. 177). Cambridge, MA: Harvard University Press.

© 2019 by DNG Econometrics, Inc

[10] On this, see Box, G. E. P., &amp; Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, B26,* 211–243.

[11] We say in general because it is quite possible that these equations are not independent of each other. In Chapter 2, we will discuss this problem more carefully.

© 2019 by DNG Econometrics, Inc

https://doi.org/10.4135/9781071802571

Appendix A: Basics of Matrix Algebra

Appendix A: Basics of Matrix Algebra

In what follows, we will denote matrices by bold capital letters and vectors by bold lowercase letters.

A.1 Definitions

A.1.1 Matrix

A matrix is a rectangular array of real numbers arranged in rows and columns. More formally, a matrix of **order**, or **dimension**, *m* by *n* (written as *m* × *n*) is a set of *m* × *n* real numbers arranged in *m* rows and *n* columns, as the matrix ***A*** below.

(A.1)

$$A_{m \times n} = [a_{ij}] = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$$

where aij is the element appearing in the *i*th row and *j*th column and where [aij] is a shorthand expression for the matrix ***A*** whose typical element is aij. The subscript on the matrix ***A*** indicates the dimension of the matrix, that is, the number of rows and columns, *m* and *n*, respectively, in this case. Of course, in practice *m* and *n* will have numerical values. Thus, a matrix $\boldsymbol{B}2\times3$ means this matrix has 2 rows and 3 columns.

A.1.2 Scalar

A scalar is single real number. Alternatively, a scalar is a 1 × 1 matrix.

A.1.3 Column Vector

A matrix consisting of *m* rows and one column (*m* × 1) is called a **column vector**. Thus,

© 2019 by DNG Econometrics, Inc

(A.2)

$$x_{3 \times 1} = \begin{bmatrix} 3 \\ 4 \\ 7 \end{bmatrix}$$

is an example of a 3 × 1 column vector.

A.1.4 Row Vector

A matrix consisting of only 1 row and *n* columns is called a **row vector**. Thus,

(A.3)

$$x_{1 \times 4} = \begin{bmatrix} 1 & 3 & 8 & -4 \end{bmatrix}$$

is an example of a 1 × 4 row vector.

A.1.5 Transposition of a Matrix

The transpose of an *m* × *n* matrix ***A***, denoted by ***A′*** (read as ***A*** prime or ***A*** transpose) is an *n* × *m* matrix obtained by interchanging the rows and columns of ***A***, that is, the *i*th row of ***A*** becomes the *i*th column of ***A′***. Thus,

(A.4)

$$A_{4 \times 2} = \begin{bmatrix} 5 & 9 \\ 3 & 11 \\ 8 & 2 \\ -7 & 6 \end{bmatrix} \quad A'_{2 \times 4} = \begin{bmatrix} 5 & 3 & 8 & -7 \\ 9 & 11 & 2 & 6 \end{bmatrix}$$

As you can see, the transpose of a matrix is obtained by rewriting its columns as rows. Note that some textbooks use $\boldsymbol{A}^{\mathrm{T}}$ to denote the transpose of ***A***.

A.1.6 Transpose of Vectors

The transpose of a row vector is a column vector, and vice versa. Usually, a vector with a prime sign (′) or a transpose sign (T) is a row vector and a vector without such a sign is a column vector.

A.1.7 Submatrix

Given an $m \times n$ matrix ***A***, if we delete all but r rows and s columns, the resulting $r \times s$ matrix is called a submatrix of ***A***. Thus, if

$$A_{3 \times 3} = \begin{bmatrix} 5 & 6 & 7 \\ 10 & 2 & 3 \\ 4 & 8 & 9 \end{bmatrix}$$

and if we delete the third row and the third column of this matrix, we obtain

(A.5)

$$B_{2 \times 2} = \begin{bmatrix} 5 & 6 \\ 10 & 2 \end{bmatrix}$$

which is a submatrix of ***A*** of order 2 × 2.

A.2 Types of Matrices

A.2.1 Square Matrix

A matrix that has the same number of rows as columns is called a **square matrix**. The ***A*** and ***B*** matrices above are examples of square matrices.

A.2.2 Diagonal Matrix

A square matrix with at least one nonzero element on the main diagonal (running from the upper-left corner to the lower right-hand corner) and zero elsewhere is called a **diagonal matrix**. Another way of expressing this is that a square matrix with off-diagonal elements all zero is called a diagonal matrix. An example is

(A.6)

$$A_{3 \times 3} = \begin{bmatrix} -4 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

© 2019 by DNG Econometrics, Inc

A.2.3 Identity or Unit Matrix

A diagonal matrix whose diagonal elements are all 1 is called an **identity**, or **unit**, **matrix** and is denoted by ***I***. The following is an example of an identity matrix.

(A.7)

$$I_{3\times3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

A.2.4 Scalar Matrix

A diagonal matrix whose diagonal elements are equal is called a **scalar matrix**. An example is the variance–covariance matrix of the error term ***u*** in the classical linear regression model, namely,

(A.8)

$$\text{var-cov}(\boldsymbol{u}) = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

As you can see, an identity matrix is a special kind of a scalar matrix.

A.2.5 Symmetric Matrix

If a square matrix ***A*** is the same as its transpose, it is called a symmetric matrix, that is, if $\boldsymbol{A}' = \boldsymbol{A}$. In this case, the element a_{ij} of matrix ***A*** is equal to the element a_{ji} of $\boldsymbol{A}'$. An example is the variance–covariance matrix of the error term ***u*** in the classical linear regression model. See, for example, Equation (A.1).

A.2.6 Null Matrix

A matrix whose elements are all zero is called a null matrix and is denoted by **0**.

A.2.7 Null Vector

A row or column vector whose elements are all zero is called a **null vector** and is also denoted by **0**.

A.2.8 Equal Matrices

Two matrices ***A*** and ***B*** are said to be equal if they are of the same order and their corresponding elements are equal, that is, $a_{ij} = b_{ij}$ for all *i* and *j*. As an example,

(A.9)

$$A_{3 \times 3} = \begin{bmatrix} 5 & 7 & 0 \\ 0 & -2 & 3 \\ 4 & 1 & 8 \end{bmatrix} \text{ and } B_{3 \times 3} = \begin{bmatrix} 5 & 7 & 0 \\ 0 & -2 & 3 \\ 4 & 1 & 8 \end{bmatrix}$$

are equal, because ***A*** = ***B***.

A.2.9 Upper Triangular Matrix

A square matrix with zeros below the main diagonal.

A.2.10 Lower Triangular Matrix

A square matrix with zeros above the main diagonal.

A.2.11 Idempotent Matrix

A square matrix ***A*** such that $\boldsymbol{A}^n = \boldsymbol{A}$ is called an **idempotent (of the same power) matrix**. That is, no matter how many times you multiply the matrix by itself, the resulting product matrix is the same as the original matrix ***A***. Thus, $\boldsymbol{A}^2 = \boldsymbol{A}$, $\boldsymbol{A}^3 = \boldsymbol{A}$, and so on. This is because all the eigenvalues of such a matrix are 1 or 0. Note that $\boldsymbol{A}^0 = \boldsymbol{I}$. On eigenvalues, see Section A.11.

A.3 Matrix Operations

A.3.1 Matrix Addition

Let ***A*** = [*aij*] and ***B*** = [*bij*]. If both of these matrices are of the same order, we define matrix addition as

(A.10)

$$A + B = C$$

where ***C*** is of the same order as ***A*** and ***B*** and is obtained by taking the sums *cij* = *aij* + *bij* for all *i* and *j*, that is, ***C*** is obtained by adding the corresponding elements of ***A*** and ***B***. If this can be done, ***A*** and ***B*** are said to be **conformable** for addition. For instance,

$$A = \begin{bmatrix} 3 & -2 & 4 & 6 \\ 8 & 7 & 4 & 11 \end{bmatrix} \text{and } B = \begin{bmatrix} 5 & 8 & 4 & -3 \\ 9 & 1 & 7 & 6 \end{bmatrix}$$

$$C = A + B = \begin{bmatrix} 8 & 6 & 8 & 3 \\ 17 & 8 & 11 & 17 \end{bmatrix}$$

A.3.2 Matrix Subtraction

Matrix subtraction follows the same principle as matrix addition except that ***C*** = ***A*** − ***B***. For example, using the matrices ***A*** and ***B*** given above, we find that

$$C = A - B = \begin{bmatrix} -2 & -10 & 0 & 9 \\ -1 & 6 & -3 & 5 \end{bmatrix}$$

A.3.3 Scalar Multiplication

To multiply a matrix ***A*** by a scalar *c* (a real number), we multiply each element of the matrix by *c*. Therefore, *c**A*** = ***A****c* = [*caij*]. Thus, if *c* = 2 and

$$A = \begin{bmatrix} -8 & 7 \\ 6 & 5 \end{bmatrix},$$

then

© 2019 by DNG Econometrics, Inc

$$2A = \begin{bmatrix} -16 & 14 \\ 12 & 10 \end{bmatrix}$$

A.3.4 Matrix Multiplication

Let ***A*** be *m* × *n* and ***B*** be *n* × *p*. Then the product ***AB***, in that order, is defined to be a new matrix ***C*** of order *m* × *p*, such that

(A.11)

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}, \quad i = 1, 2, \ldots, m \text{ and } j = 1, 2, \ldots, p$$

That is, the element in the *i*th row and the *jth* column of ***C*** is obtained by multiplying the elements of the *i*th row of ***A*** by the corresponding elements of the *j*th column of ***B*** and summing over terms. This is known as the **row by column rule of multiplication**. Thus, to obtain c_{11}, the element in the first row and first column of ***C***, we multiply the elements in the first row of ***A*** by the corresponding elements in the first column of ***B*** and sum over all terms, and so on.

Note that for multiplication to exist, matrices ***A*** and ***B****must be conformable with respect to multiplication*, that is, the number of columns in ***A*** must be equal to the rows in ***B***. For example,

$$A_{2 \times 3} = \begin{bmatrix} 3 & 4 & 7 \\ 5 & 6 & 1 \end{bmatrix} \text{and } B_{3 \times 2}\begin{bmatrix} 2 & 1 \\ 3 & 5 \\ 6 & 2 \end{bmatrix}$$

$$AB = C_{2 \times 2} = \begin{bmatrix} (3 \times 2) + (4 \times 3) + (7 \times 6) & (3 \times 1) + (4 \times 5) + (7 \times 2) \\ (5 \times 2) + (6 \times 3) + (1 \times 6) & (5 \times 1) + (6 \times 5) + (1 \times 2) \end{bmatrix}$$

$$= \begin{bmatrix} 60 & 37 \\ 34 & 37 \end{bmatrix}$$

Note that in this example the resulting ***C*** matrix is of order 2 × 2.

In this example, is the product ***BA*** defined? If so, what is the dimension of the resulting product matrix?

© 2019 by DNG Econometrics, Inc

A.3.4.1 Properties of Matrix Multiplication

1. Matrix multiplication is not necessarily **commutative**, that is, in general, ***AB*** ≠ ***BA***. The order in which matrices are multiplied is very important. ***AB*** means ***A*** is **postmultiplied** by ***B*** or ***B*** is premultiplied by ***A***.
2. Even if ***AB*** and ***BA*** exist, the resulting matrices may not be of the same order. For example, if ***A*** is $m \times n$ and ***B*** is $n \times m$, ***AB*** is $m \times m$, but ***BA*** is $n \times n$, and hence of a different order.
3. Even if ***A*** and ***B*** are both square matrices, so that ***AB*** and ***BA*** are both defined, the resulting matrices will not necessarily be equal. For example, consider

$$A = \begin{bmatrix} 4 & 7 \\ 3 & 2 \end{bmatrix} \text{and } B = \begin{bmatrix} 1 & 5 \\ 6 & 8 \end{bmatrix}; \text{ then } AB = \begin{bmatrix} 46 & 76 \\ 15 & 31 \end{bmatrix} \text{and } BA = \begin{bmatrix} 19 & 17 \\ 48 & 58 \end{bmatrix}$$

Thus, ***AB*** ≠ ***BA***. But if both ***A*** and ***B*** are **identity matrices**, then ***AB*** = ***BA***.

4. A row vector multiplied by a column vector is a scalar. As an example, consider the OLS residuals $e1$, $e2$, …, en. Letting **e** be a column vector of the residuals and **e′** the row vector of the residuals, we can see that

(A.12)

$$e'e = \begin{bmatrix} e_1 & e_2 & e_3 & \cdots & e_n \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} = e_1^2 + e_2^2 + e_3^2 + \cdots + e_n^2 = \sum_1^n e_i^2$$

which is a scalar, or 1 × 1 matrix. Note that **e′** is $1 \times n$, but **e** is $n \times 1$.

5. But see what happens if we multiply an $n \times 1$ column vector by a $1 \times n$ row vector. We get an $n \times n$ matrix:

(A.13)

$$ee' = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix} \begin{bmatrix} e_1 & e_2 & e_3 & \cdots & e_n \end{bmatrix} = \begin{pmatrix} e_1^2 & \dots & e_1e_n \\ \vdots & \ddots & \vdots \\ e_ne_1 & \cdots & e_n^2 \end{pmatrix}$$

6. A matrix multiplied by a column vector is a column vector. If $\boldsymbol{A}_{m\times n}$ and $\boldsymbol{B}_{n\times 1}$, then ***AB*** is ($m$ × 1), a column vector.
7. A row vector postmultiplied by a matrix is a row vector.
8. Matrix multiplication is **associative**, that is, (***ABC***) = ***A***(***BC***), where ***A*** is $m \times n$, ***B*** is $n \times p$, and ***C*** is $p \times k$.
9. Matrix multiplication is **distributive** with respect to addition, thus

$$\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C})=\boldsymbol{AB}+\boldsymbol{AC}\text{and}(\boldsymbol{B}+\boldsymbol{C})\boldsymbol{A}=\boldsymbol{BA}+\boldsymbol{CA}$$

A.4 Matrix Transposition

1. The transpose of a transposed matrix is the original matrix: (***A'***)′ = ***A***.
2. The transpose of the sum of two (conformable) matrices is the sum of their respective transposes. Thus, if ***A*** and ***B*** are conformable for addition, then
   $C = A + B$ and $C' = (A + B)' = A' + B'$
3. If the product ***AB*** is defined, then (***AB***)′ = ***B'A'***. This can be generalized. If the product (***ABCD***) is defined, then (***ABCD***)′ = ***D'C'B'A'***.
4. The transpose of an identity matrix ***I*** is the identity matrix itself, that is, ***I′*** = ***I***.
5. The transpose of a scalar is the scalar itself. Thus, if γ is a scalar, γ′ = *γ*.
6. The transpose of (γ***'A***)′ is γ***A'***. This is because (γ***'A***)***'*** = = ***A'***γ = γ***A'***.
7. If ***A*** is a square matrix such that ***A*** = ***A′***, then ***A*** is a symmetric matrix.

A.5 Matrix Inversion

An inverse of a square matrix ***A***, denoted by $\boldsymbol{A}^{-1}$ (read as ***A*** inverse), if it exists, is a unique square matrix such that

$$AA^{-1} = A^{-1}A = I$$

where ***I*** is an identity matrix whose order is the same as that of ***A***. For example,

$$A = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix} A^{-1} = \begin{bmatrix} -1 & \frac{1}{2} \\ \frac{6}{8} & -\frac{1}{4} \end{bmatrix} AA^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

Properties of matrix inversion are given in Section A.8.

A.6 Determinants

To every square matrix ***A***, there corresponds a number (scalar) known as the determinant of the matrix, which is denoted by det ***A*** or by the symbol |***A***|, the two parallel lines that enclose ***A*** meaning “the determinant of” and not the usual symbol for the absolute value. Note that a matrix per se has no numerical value, but the determinant of a matrix is a number.

A.6.1 Evaluation of a Determinant

The process of finding the value of a determinant is known as **evaluation**, **expansion**, or **reduction** of the determinant. This is accomplished by manipulating the entries of the matrix in a well-defined manner.

Evaluation of a 2×2 determinant:

If

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

its determinant is evaluated as

$$|A| = \begin{vmatrix} a_{11} \searrow & \swarrow a_{12} \\ a_{21} \nearrow & \nwarrow a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

which is obtained by cross-multiplying the elements on the main diagonal and subtracting from this the cross-

© 2019 by DNG Econometrics, Inc

multiplication of the elements on the other diagonal of matrix ***A***, as shown by the arrows.

Evaluation of a 3×3 determinant:

If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$|A| = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}$$

An examination of the preceding pattern is not arbitrary. Note the following:

1. Each term in the expansion of the determinant contains one and only one element from each row and each column.
2. The number of elements in each term is the same as the number of rows or columns in the matrix: A 2 × 2 determinant has two elements in each term of expansion, a 3 × 3 determinant has three elements in each term of expansion, and so on.
3. The general rule is that the determinant of order $n \times n$ has $n! = n \times (n-1) \times \ldots 3 \times 2 \times 1$ terms in the expansion where $n!$ is called "*n* factorial." Thus, a 6 × 6 determinant will have 6 × 5 × 4 × 3 × 2 × 1 = 720 terms in the expansion.
4. The terms in the expansion alternate in sign from + to −.

A.6.2 Properties of Determinants

1. A matrix whose determinant is zero is called a **singular matrix**, whereas a matrix with a nonzero determinant is called a **nonsingular matrix**. The inverse of a singular matrix does not exist.
2. If all the elements of any row of ***A*** are zero, its determinant is zero.
3. |***A'***| = |***A***|, that is, the determinants of ***A*** and ***A*** transpose are the same.
4. Interchanging any two rows or any two columns of ***A*** changes the sign of λ.
5. ;If every element of a row or a column of ***A*** is multiplied by a scalar λ then |***A***| is multiplied by λ.
6. ;If two rows or columns of a matrix are identical, its determinant is zero.
7. If one row or a column is a multiple of another row or column of that matrix, its determinant

is zero.

8. |***AB***|= |***A***| |***B***|, that is, the determinant of the product of two matrices is the product of their individual determinants.
9. If ***A*** = diag(a_1, a_2, . . . , a_n), and all off-diagonal elements are zero, |***A***|= a_1, a_2, . . . , a_n, that is, the product of all the diagonal elements.

A.7 Rank of a Matrix

The rank of a matrix is the maximum number of linearly independent rows, which is the same as the maximum number of linearly independent columns. Thus, the rank of a matrix is equal to that of its transpose. If a matrix has *m* rows and *n* columns, with $m \le n$, then the rank is $\le m$; if *m* &gt; *n*, then the rank is $\le n$. If the rank is equal to the smaller of *m* and *n*, then the matrix is of **full rank**. The rank of a matrix can also be defined as the order of the largest square submatrix whose determinant is not zero.

For example, for the matrix

$$A = \begin{bmatrix} 3 & 6 & 6 \\ 0 & 4 & 5 \\ 3 & 2 & 1 \end{bmatrix}$$

its determinant is zero because Row 2 + Row 3 = Row 1, hence it is a singular matrix. Hence, although its order is 3 × 3, its rank is less than 3. Actually it is 2, because we can find a 2 × 2 submatrix whose determinant is not zero. For instance, if we delete the first row and the first column of ***A***, we obtain

$$B = \begin{bmatrix} 4 & 5 \\ 2 & 1 \end{bmatrix}$$

The determinant of this matrix is −6, which is nonzero, thus establishing that the rank of the matrix ***A*** is in fact 2.

A.7.1 Minor

If the *i*th row and the *j*th column of an *n* × *n* matrix ***A*** are deleted, the determinant of the resulting submatrix is called the **minor** of the element a_{ij}, the element at the intersection of the *i*th row and *j*th column, and is denoted by $|\boldsymbol{M}_{ij}|$.

As an example, consider the following matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

The minor of a_{11} is

$$|M_{11}| = \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix} = a_{22}a_{33} - a_{23}a_{32}.$$

The minors of other elements of $\boldsymbol{A}$ can be found similarly.

A.7.2 Cofactor

The cofactor of the element a_{ij} of an $n \times n$ matrix $\boldsymbol{A}$, denoted by c_{ij}, is defined as $c_{ij} = (-1)^{i+j}|M_{ij}|$. In words, a cofactor is a **signed minor**, the sign being positive if $i + j$ is even and being negative if $i + j$ is odd. For instance, the cofactor of the element a_{11} of a 3 × 3 matrix $\boldsymbol{A}$ given previously is $a_{22}a_{33} - a_{23}a_{32}$, whereas the cofactor of the element a_{21} is $–(a_{12}a_{33} – a_{13}a_{32})$ since the sum of the subscripts 2 and 1 is 3, which is an odd number.

A.7.3 Cofactor Matrix

Replacing the elements of a_{ij} of a matrix $\boldsymbol{A}$ by their cofactors, we obtain a matrix known as the **cofactor matrix** of $\boldsymbol{A}$, denoted by (cof $\boldsymbol{A}$).

A.7.4 Adjoint Matrix

The adjoint matrix, written as (adj $\boldsymbol{A}$), is the transpose of the cofactor matrix, that is, (adj $\boldsymbol{A}$) = (cof $\boldsymbol{A}$)′.

A.8 Finding the Inverse of a Square Matrix

If ***A*** is square and nonsingular, that is, $|\boldsymbol{A}| \neq 0$, its inverse can be found as follows:

$$A^{-1} = \frac{1}{|A|}(\text{adj } A)$$

The steps involved in the computation of the inverse are as follows:

1. Find the determinant of ***A***; if it is nonzero, proceed to Step 2.
2. Replace element *aij* of ***A*** by its cofactors to obtain the cofactor matrix.
3. Transpose the cofactor matrix to obtain the adjoint matrix.
4. Divide each element of the adjoint matrix by $|\boldsymbol{A}|$.
5. If you multiply ***A*** by $\boldsymbol{A}^{-1}$, you should get an identity matrix.

Example: Consider the following matrix:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 7 & 4 \\ 2 & 1 & 3 \end{bmatrix}$$

It is left for the reader to verify that the inverse of this matrix is as follows:

$$A^{-1} = -\frac{1}{24}\begin{bmatrix} 17 & -3 & -13 \\ -7 & -3 & 11 \\ -9 & 3 & -3 \end{bmatrix}$$

A.8.1 Properties of Inverse Matrices

1. If ***A*** and ***B*** are square matrices of the same size such that both are nonsingular, then $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$.
2. If ***A*** is nonsingular, then $(\boldsymbol{A}')^{-1} = (\boldsymbol{A}^{-1})'$, that is, the inverse of the transposed matrix is the transpose of its inverse. That is, the inverse symbol (−1) and the transpose symbol (′) are interchangeable.
3. If ***A*** is nonsingular,
   $$\left|\boldsymbol{A}^{-1}\right| = \frac{1}{|\boldsymbol{A}|}.$$
4.

$(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$. That is, the inverse of a nonsingular matrix is the original matrix.

5. If $\boldsymbol{A}$ is nonsingular, the system of equation $\boldsymbol{Ax} = \boldsymbol{c}$ has the unique solution $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{c}$, where $\boldsymbol{x}$ and $\boldsymbol{c}$ are appropriately defined vectors.

A.9 Trace of a Square Matrix

The trace of a square matrix $\boldsymbol{A}$, denoted by tr($\boldsymbol{A}$), is defined as the sum of its diagonal elements, that is,

$$\mathrm{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn}$$

Some of the properties of the trace are as follows:

1. tr(k) = k, where k is a constant.
2. tr($\boldsymbol{A'}$) = tr($\boldsymbol{A}$), that is, the trace of a transposed matrix is equal to the trace of the original matrix.

A.10 Quadratic Forms and Definite Matrices

If $\boldsymbol{A}$ is a symmetric matrix and $\boldsymbol{x}$ is a vector, the product

$$x'Ax = \sum_{i}^{n}\sum_{j}^{n} a_{ij}x_ix_j$$

is called a **quadratic form**.

Note that the symmetry of $\boldsymbol{A}$ means
$a_{ij}x_ix_j = a_{ji}x_ix_j$.

A quadratic form is called **positive definite (PD)** if $\boldsymbol{x'Ax}$ &gt; **0** for all non-null vectors. It is called **positive semidefinite (PSD)** if $\boldsymbol{x'Ax} \geq \boldsymbol{0}$. It is called **negative definite (ND)** if $\boldsymbol{x'Ax}$ &lt; **0** and **negative semidefinite (NSD)** if $\boldsymbol{x'Ax} \leq \boldsymbol{0}$. These matrices play an important role in statistics. Note that *a quadratic form is a scalar*, hence it equals its trace, that is, $\boldsymbol{x'Ax}$ = tr($\boldsymbol{x'Ax}$) = tr($\mathbf{A}\boldsymbol{xx'}$).

© 2019 by DNG Econometrics, Inc

A.10.1 Some Properties of Quadratic Forms

1. If ***A*** is an $n \times n$ PD matrix, then $|\boldsymbol{A}|$ &gt; 0, rank of ***A*** = *n* and ***A*** is nonsingular.
2. If ***A*** is PD, then $\boldsymbol{A}^{-1}$ is also PD.
3. If ***A*** is PD, there exists a nonsingular matrix ***P*** such that $\boldsymbol{PAP'} = \boldsymbol{I}$ and $\boldsymbol{P'P} = \boldsymbol{A}^{-1}$.
4. If ***A*** and ***B*** are symmetric and $(\boldsymbol{A} - \boldsymbol{B})$ is PD, then $|\boldsymbol{A}| \geq |\boldsymbol{B}|$ and $\boldsymbol{x'Ax} \geq \boldsymbol{x'Bx}$ for all ***x***.

A.10.2 Mean and Variance of Quadratic Forms[1]

If ***x*** is a random vector with mean μ and covariance Σ and if ***A*** is a symmetric matrix of constants, then

$$E(x'Ax) = \mathrm{tr}(A\Sigma) + \mu' A\mu$$

and its variance is

$$\mathrm{var}(x'Ax) = 2\mathrm{tr}\left[(\boldsymbol{A}\Sigma)^2\right] + 4\mu' \boldsymbol{A}\Sigma\boldsymbol{A}\mu$$

If

$X \sim N(0, \Sigma),\ \textit{then}\ \mathrm{var}(x'Ax) = 2\mathrm{tr}(A\Sigma)^2.$

A.11 Eigenvalues and Eigenvectors

For every square matrix ***A***, we can find a scalar λ and a nonzero vector ***x*** such that

(A.14)
$$Ax = \lambda x$$

In this equation, λ is called an **eigenvalue**, also known as a **characteristic root**, and ***x*** is known as an **eigenvector**, also known as a **characteristic vector**.

To find λ and ***x*** of ***A***, we can write (A.14) as

(A.15)
$$(A - \lambda I)x = \mathbf{0}$$

If the matrix $(\boldsymbol{A} - \lambda\boldsymbol{I})$ is nonsingular, the only solution to (A.15) is $\boldsymbol{x} = \mathbf{0}$. However, a nonzero solution is possible if $(\boldsymbol{A} - \lambda\boldsymbol{I})$ is singular, which means the determinant $|(\boldsymbol{A} - \lambda\boldsymbol{I})| = \mathbf{0}$. In this situation, we obtain a

© 2019 by DNG Econometrics, Inc

polynomial of order *n* in which is known as the **characteristic polynomial** in ***A*** and the resulting equation is called the **characteristic equation**. The roots of the characteristic equation are called characteristic roots or eigenvalues. Corresponding to each λ_i we will have a characteristic vector $\boldsymbol{x_i}$. In all, there will be *n* characteristic roots and characteristic vectors. It may be noted that not all characteristic roots will be distinct.

Example: Consider the following matrix

(A.16)

$$A = \begin{pmatrix} 2 & 2 \\ 2 & -1 \end{pmatrix}$$

Using this matrix and (A.15), we obtain the following determinant:

(A.17)

$$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 2 \\ 2 & -1-\lambda \end{vmatrix} = \lambda^2 - \lambda - 6 = 0$$

Since this is a polynomial of second degree, we will obtain two characteristic roots, namely, $\lambda_1 = 3$ and $\lambda_2 = –2$.

Using the first root and (A.14), we obtain

(A.18)

$$\begin{bmatrix} 2 & -3 & 2 & \\ 2 & -1 & & -3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -1 & 2 \\ 2 & -4 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Since the two rows of this matrix are linearly dependent, there is an infinite number of solutions, which can be expressed by the equation $x_1 = 2x_2$. To get a unique solution, we **normalize** the solution by imposing the restriction that $x_1^2 + x_2^2 = 1,$ which yields $x_1^2 + x_2^2 = (2x_2)^2 + x_2^2 = 5x_2^2 = 1$. Taking the positive square root of this, we obtain $x_2 = \frac{1}{\sqrt{5}}$. Since

$$x_1 = 2x_2 = \frac{2}{\sqrt{5}},$$

the first characteristic vector is $\boldsymbol{v}_1 = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \end{bmatrix}$. The reader can verify that corresponding to the eigenvalue of −2, we obtain the second characteristic vector as$\boldsymbol{v}_2 = \begin{bmatrix} -1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$.

Note that

$\boldsymbol{v}'_1\boldsymbol{v}_1 = 1$ *and* $\boldsymbol{v}'_1\boldsymbol{v}_2 = 0$.

These two important properties can be generalized to as many as *n* characteristic vectors. When the scalar product of two vectors is zero, we call them **orthogonal vectors**, that is, they are perpendicular to each other. The other property of the scalar product of a characteristic root with itself is unity is because of the normalization rule. As a result of these two properties, the characteristic vectors of a matrix form what is called a set of **orthonormal vectors**.

The knowledge of characteristic roots and characteristic vectors is quite useful in determining the properties of quadratic forms, more specifically the following.

1. The quadratic form ***x'Ax*** is positive (negative) definite, if and only if *every* characteristic root of ***A*** is positive (negative).
2. ***x'Ax*** is positive (negative) semidefinite, if and only if all *characteristic roots* of ***A*** are nonnegative (nonpositive) and *at least* one root is zero.
3. ***x'Ax*** is indefinite, if and only if some of the characteristic roots of ***A*** are positive and some are negative.

The knowledge of characteristic roots and vectors is also useful in the study of symmetric matrices. Some examples are as follows:

1. For any symmetric matrix ***A***, there exists an *orthogonal* matrix ***H*** such that ***H'AH*** = ***Z***, where ***Z*** is a diagonal matrix whose diagonal elements are the characteristic roots of ***A***. Note that ***H'H*** = ***I***.
2. The rank of a square matrix is equal to the number of its nonzero characteristic roots.
3. For any matrices ***X*** and ***Y***, not necessarily square, the nonzero characteristic roots of ***XY*** and ***YX*** are the same, whenever both these product matrices are defined.
4. If ***A*** and ***B*** are symmetric matrices of the same size, then they both can be diagonalized by the same orthogonal matrix if and only if ***AB*** = ***BA***.
5. The determinant of a square matrix is the product of its characteristic roots.
6. The trace of a square matrix is the sum of its characteristic roots.

A.12 Vector and Matrix Differentiation

If $\boldsymbol{y} = f(\boldsymbol{x})$, a function of the variables $x1,x2,\ldots,xk$, in $\boldsymbol{x} = (x1,x2,\ldots,xk)'$, and let
$\partial y / \partial x_1,\ \partial y / \partial x_2,\ \ldots,\ \partial y / \partial x_k$
be the partial derivatives, then we define $\partial y / \partial x$ as

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{bmatrix} \partial y / \partial x_1 \\ \partial y / \partial x_2 \\ \vdots \\ \partial y / \partial x_k \end{bmatrix}$$

which we can also write as

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}'} = \left(\frac{\partial y}{\partial x_1}, \frac{\partial y}{\partial x_2}, \cdots, \frac{\partial y}{\partial x_n} \right)$$

which is the transpose of
$\partial \boldsymbol{y} / \partial \boldsymbol{x}$.

If
$a' = \begin{bmatrix} a_1 & a_2 & \cdots & a_k \end{bmatrix}$
is a row vector of numbers and

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

is a column vector of the variables $x1,x2,\ldots,xk$,, and if $\boldsymbol{y} = \boldsymbol{a'x} = \boldsymbol{x'a}$, then

$$\frac{\partial \boldsymbol{a}'\boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}'\boldsymbol{a}}{\partial \boldsymbol{x}} = a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{bmatrix}$$

A.12.1 Some Differentiation Rules

Let ***A*** be a matrix and ***a***, ***x***, ***y*** be vectors. Assuming the usual matrix rules about multiplication, division (inverse), and the like, the following rules exist:

1. $$\frac{\partial y' x}{\partial x} = y$$
2. $$\frac{\partial x' A x}{\partial x} = (A + A')x$$
$$= 2Ax = 2A'x \text{ ifAis symmetric}$$

3. $\frac{\partial Ax}{\partial x} = A'$
4. $\frac{\partial Ax}{\partial x'} = A$

A.12.2 The Hessian Matrix

Let $\boldsymbol{x} = (x1,x2,\ldots,xk)'$ be an $n \times 1$ vector and $f(\boldsymbol{x})$ a real function differentiable with respect to x_i, an element of the $\boldsymbol{x}$ vector. Let

$m(x) = \frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}}$

represent first derivatives. Let $H(\boldsymbol{x})$ represent the first derivative of $m(\boldsymbol{x})$, that is, the second derivative of the original function $f(\boldsymbol{x})$ as follows:

$$\boldsymbol{H}(\boldsymbol{x}) = \frac{\partial m(\boldsymbol{x})}{\partial \boldsymbol{x}'} = \begin{bmatrix} \frac{\partial m_1(x)}{\partial x_1} & \cdots & \frac{\partial m_1(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial m_n(x)}{\partial x_1} & \cdots & \frac{\partial m_n(x)}{\partial x_n} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}$$

$\boldsymbol{H}(\boldsymbol{x})$ is called the Hessian matrix or just the Hessian. This matrix is very useful in determining the local extreme (minimum or maximum) value of f. A necessary condition for $\boldsymbol{x} = \boldsymbol{x}0$ being a local extreme of f is

$m(x_0) = \boldsymbol{0}$

If this condition holds, then

if $\boldsymbol{H}(\boldsymbol{x}0)$ is positive definite, $\boldsymbol{x}0$ is a local minimum.
if $\boldsymbol{H}(\boldsymbol{x}0)$ is negative definite, $\boldsymbol{x}0$ is a local maximum.

Notes

[1] For proofs, see Searle, S. R. (1971). *Linear models* (pp. 55–57). New York, NY: Wiley.

Appendix B: Essentials of Large-Sample Theory1

Appendix B: Essentials of Large-Sample Theory[1]

The behavior of a statistic in small, or finite, samples and in large, or infinite, samples is not always the same. For example, in the classical linear regression model, we showed that the estimated error variance obtained by ordinary least squares is unbiased, whether in small or large samples, but that obtained by the method of maximum likelihood is biased in small samples but is unbiased as the sample size increases indefinitely.

In many cases, it is not easy to estimate the statistical properties of an estimator in finite samples, but it may be possible to do so in large, or infinite, samples. For example, we can derive the sampling distribution of the sample mean, $\bar{X}$, but it is not easy to derive the sampling distribution of $1 \big/ \bar{X}$, which is a nonlinear function of the sample mean.

In this appendix, we consider several aspects of **large-sample theory**, also known as **asymptotic theory**, the term *asymptotic* meaning the limiting behavior of a variable or a statistic of interest as the sample size increases indefinitely. In what follows, we discuss some salient aspects of asymptotic theory.

To develop this theory, we need some background, which is provided by some well-known inequalities and theorems.

B.1 Some Inequalities

B.1.1 Markov's Inequality

Let *X* be a nonnegative random variable, and suppose that *E*(*X*) exists. Then for any *a* &gt; 0,

(B.1)

Pr(*X* &gt; *a*) ≤ $\frac{E(X)}{a}$

Assuming *X* is a continuous random variable with density function *f*(*X*),

$$E(X) = \int_0^\infty xf(x)\mathrm{d}x$$
$$= \int_0^a xf(x)\mathrm{d}x + \int_a^\infty xf(x)\mathrm{d}x$$
$$\geq \int_a^\infty xf(x)\mathrm{d}x$$
$$\geq \int_a^\infty af(x)\mathrm{d}x$$
$$\geq a\int_a^\infty f(x)\mathrm{d}x$$
$$= a\Pr(X \geq a)$$

or

(B.2)

$$\frac{E(X)}{a} \geq \Pr(X \geq a)$$

which is the required result. Essentially, Markov's inequality states that the probability that *X* is much bigger than *E*(*X*) is small.

A great advantage of Markov's inequality is that it enables us to place an upper bound on the probability that *g*(*X*) ≥ *a* as long as *g*(*X*) is a nonnegative value and *Eg*(*X*) exists.

B.1.2 Chebyshev's Inequality

As a corollary of Markov's inequality, we obtain Chebyshev's inequality. If *X* is a random variable with mean μ and variance σ^2, then for any value *k* &gt; 0, it can be shown that

(B.3)[2]

$$\Pr(\left| X - \mu \right| \geq k\sigma) \leq \frac{1}{k^2};\ k \text{ &gt; } 0$$

This inequality is important because it provides a bound for all random variables. For example, it shows that the probability that the value of a random variable deviates more than three standard deviations (*k* = 3) from its mean value is less than 1/9.

Proof: We can use Markov's inequality to prove Chebyshev's inequality. Letting $a = (k\sigma)^2$ and noting that $(X–\mu)^2$ is a nonnegative random variable, we have

(B.4)

$$[\Pr(X - \mu)^2 \geq k^2] \leq \frac{E[(X - \mu)^2]}{k^2}$$

Since

$(X - \mu)^2 \geq k^2\sigma^2$ if and only if $\left| X - \mu \right| \geq k\sigma$,

Equation (B.4) is equivalent to

(B.5)

$$\Pr\left\{ \left| X - \mu \right| \geq k\sigma \right\} \leq \frac{E[(X - \mu)]^2}{k^2\sigma^2} = \frac{1}{k^2}$$

which is the required result.

The reason Markov’s and Chebyshev’s inequalities are important is because they enable us to establish bounds on probabilities just knowing only the mean or both the mean and variance of the probability distribution without knowing the actual probability distribution. If the latter were known, there would be no need to look for the bounds as the desired probabilities could be computed exactly.

B.1.3 Khinchine’s Theorem (Weak Law of Large Numbers)

If $X_1, X_2, \ldots, X_n$ are iid random variables, with mean *μ,* then for any ε &gt;0, the sample mean

$\bar{X}_n$

converges in probability to μ as $n \to \infty$

That is,

$$\lim_{n \to \infty} \Pr \left| \bar{X}_n - \mu \right| \text{ &gt; } \varepsilon \to 0$$

Alternatively,

(B.6)

$$\lim_{n \to \infty} \Pr \left| \bar{X}_n - \mu \right| \leq \varepsilon = 1$$

In short,

$p\lim(\bar{X}_n) = \mu$,

where *p*lim stands for probability limit.

Khinchine's theorem is also known as the **weak law of large numbers** (WLLN) to distinguish it from the **strong law of large numbers** (SLLN), which we will discuss shortly.

The WLLN is easy to prove. We know that $E(\overline{X}) = E[(X_1 + X_2 + \cdots + X_n) \big/ n] = n\mu \big/ n = \mu$ and $\text{var}(\overline{X}) = \sigma^2 \big/ n$. Therefore,

$$\lim_{n \to \infty} \frac{\sigma^2}{n} \rightarrow 0$$

We can establish this theorem, using Chebyshev's inequality:

(B.7)

$$\Pr[\left| \overline{X} - \mu \right| \text{ \&gt; } \varepsilon] \leq \frac{\sigma^2}{n\varepsilon^2}; \quad \varepsilon \text{ \&gt; } 0$$

B.1.4 Strong Law of Large Numbers

The WLLN is based on the concept of the *convergence in probability* limit, whereas the SLLN is based on the concept of almost sure convergence. These concepts of convergence are discussed below.

B.1.5 Jensen's Inequality

If $Y=g(X)$ is a concave function so that it lies everywhere below its tangent line and $E(X) = \mu$, then

(B.8)

$$E(Y) \leq g(\mu)$$

It is well-known that the logarithmic function is concave, and so

(B.9)

$$E\,[\log\,(X)] \;\leq \log\,[E(X)]$$

Similarly, if $Y=g(X)$ is a convex function so that it lies everywhere above its tangent line, then

(B.10)

$$E(Y) \geq g(\mu)$$

Thus, the square function is convex. So we have

(B.11)

$$E(X^2) \geq [E(X)]^2$$

Both (B.9) and (B.10) hold regardless of the distribution of *X*.

B.1.6 Cauchy–Schwarz Inequality

If the random variables *X* and *Y* have finite variances, then

(B.12)

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

with equality if and only if Pr($aX + bY$) = 1 for some real constants *a* and *b*, at least one of which is nonzero. This inequality is used to establish that the correlation coefficient between two random variables lies between −1 and +1.

Without loss of generality, suppose $E(\boldsymbol{X})=E(\boldsymbol{Y})= 0$.

In this case,

$$E(\boldsymbol{XY}) = \text{cov}(\boldsymbol{X}, \boldsymbol{Y}) = \sigma_{XY};\ E(\boldsymbol{X}^2) = \text{var}(\boldsymbol{X});\ E(\boldsymbol{Y}^2) = \text{var}(\boldsymbol{Y})$$

and (B.12) reduces to

$$0 \leq \frac{\sigma^2_{\boldsymbol{XY}}}{\sigma^2_{\boldsymbol{X}}\sigma^2_{\boldsymbol{Y}}} \leq 1$$

$$0 \leq \rho^2 \leq 1$$

where ρ is the population coefficient of correlation defined as

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

B.1.7 The Central Limit Theorem

One of the most remarkable results in probability theory is the **central limit theorem (CLT)**, proposed by the French mathematician Laplace. In simple terms, the CLT asserts that the sum of a large number of independent variables has a (probability) distribution that is approximately normal.

Formally, let $X_1, X_2, \ldots, X_n$ be a sequence of iid random variables each with mean μ and variance σ^2. Let

$$\overline{X}_n = \Sigma_{i=1}^{n} X_i \big/ n,$$

which is the sample mean of the random variable X. Then as n increases indefinitely (ie., $n \to \infty$),

(B.13)

$$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

where ≈ means approximately. Note that this result holds true regardless of the form of the probability distribution function.

As a result, the following variable, Z, follows the standard normal distribution, namely,

(B.14)

$$Z = \frac{\overline{X}_n - \mu}{\sigma / \sqrt{n}} \approx N(0, 1)$$

That is, Z follows the standard normal distribution. Notice that Equations (B.13), (B.14), and the following equations are all equivalent:

(B.15)

$$(\overline{X} - \mu) \approx N(0, \sigma^2 \big/ n)$$

$$\sqrt{n}(\overline{X}_n - \mu) \approx N(0, \sigma^2)$$

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \approx N(0, 1)$$

We will not provide a proof of this theorem as it involves *moment-generating functions* or *characteristic functions.*[3]

Now we turn to a discussion of convergence. This topic is often discussed under the title of **limit theorems**. The general idea behind the limit theorems is to study the behavior of a stochastic quantity as the sample size n increases indefinitely.

B.2 Types of Convergence

B.2.1 Convergence of a Sequence of Real Numbers

A sequence of real numbers $\{a_n\}$, $n = 1, 2, 3, \ldots$, converges to a real number a if for any ε &gt; 0, there exists an integer N such that for all n &gt; N, we have

(B.16)
$\left|a_n - a\right|$ &lt; ε

which means as $n \rightarrow \infty$, $\lim\limits_{n \to \infty} a_n = a$.

As an example, let $a_t = 6 + (0.4)^t$, as $t \to \infty$, $a_t \to 6$. Hence, $\lim\limits_{t \to \infty} a_t = 6$. For another example, $\lim\limits_{n \to \infty} \left[1 + \frac{1}{n}\right] = 1$.

Consider now $a_t = (-0.999)^t$. As t tends to infinity, this sequence tends to 0, but in an oscillating manner.

B.2.2 Strong or Almost Sure Convergence

If for all ε &gt; 0, *then* $\Pr(\lim\limits_{n \to \infty} X_n - X) = 1$, X_n converges to X *almost surely* or with probability 1 to X, written as

$X_n \overset{\text{as}}{\rightarrow} X$,

where “as” = almost surely. Almost sure convergence is used in advanced analysis, but convergence in probability, discussed below, is easier to work with and is sufficient for most purposes.

B.2.3 Convergence in Probability of a Sequence of Random Variables

Consider a sequence of random variables $X_1, X_2, \ldots, X_n$ with (cumulative) distribution functions $F_1(\ ), F_2(\ ), \ldots, F_n(\ )$, the index n indicating the number of terms in the sequence.

This sequence is said to converge in probability to a constant c, if

(B.17)

© 2019 by DNG Econometrics, Inc

$$\lim_{n \to \infty} \Pr[\left| X_n - c \right| \text{ \&gt; } \varepsilon] = 0 \text{ for any } \varepsilon \text{ \&gt; } o \Big|$$

What (B.17) states is that the sequence of random variables $X1$, $X2$, . . . , Xn converges in probability to the constant c if the limit of this sequence of probabilities is zero for any positive value of ε.

To account for the probabilistic nature of random variables, we can modify the definition given in (B.17) as follows: A sequence of random variables {Xn}, n = 1, 2, 3, . . . is said to converge to a random variable X in *probability*, denoted as

$$X_n \xrightarrow{P} X$$

if for any ε &gt; 0, as $n \to \infty$

(B.18)

$$\Pr[\left| X_n - X \right| \geq \varepsilon] \to 0$$

or equivalently, for any

(B.19)

$$\Pr[\left| X_n - X \right| \text{ \&lt; } \varepsilon] \to 1$$

In simple terms, this means that the difference between Xn and X is likely to be small as the sample size n increases indefinitely. This concept plays a key role in determining the consistency of an estimator and the various laws of large numbers.

Equations (B.18) and (B.19) are often expressed as

(B.20)

$$p \lim_{n \to \infty} X_n = X$$

where *p*lim means probability limit.

Note that *p*lim( ) is an operator, like *E*( ), the expectations operator. But the *p*lim has some properties that are not shared by the expectations operator, which makes it easy to use *p*lim in situations where it is difficult or impossible to obtain results based on the expectations operator, as the following discussion will show. One reason for this is that the *E*( ) is a **linear operator**.

The *p*lim has the following properties, often attributed to Slutsky, a Russian mathematician. If $X_n \xrightarrow{P} X \text{ and } Y_n \xrightarrow{P} Y, \text{ then}$

1. $aX_n \xrightarrow{P} aX$ (a, real number).
2. If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.
3. If $X_n \overset{P}{\to} X$ and $Y_n \overset{P}{\to} Y$, then $X_nY_n \overset{P}{\to} XY$.
4. If $X_n \overset{P}{\to} X$ and $Y_n \overset{P}{\to} Y$, then $\frac{X_n}{Y_n} \overset{P}{\to} \frac{X}{Y}$, if Y is not 0.
5. If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.
6. If $h(\ )$ is a continuous function, then X_n ‣ X implies that $h(X_n)$ ‣$h(X)$, that is, convergence in probability is preserved under continuous transformation. Notice that properties (3) and (4) do not hold for the expectations operator. Thus, $E(XY)$ is not equal to $E(X)\cdot E(Y)$, unless X and Y are independent. Similarly, $E(X/Y)$ is *not* equal to $E(X)/E(Y)$.

Besides convergence in probability, there are two other modes of convergence: *convergence in mean square* and *convergence in distribution*.

B.2.4 Convergence in Mean Square

A sequence $\{X_n\}$, n = 1, 2, 3, … is said to converge to X in *mean square* if

(B.21)

$$\lim_{n \to \infty} E(X_n - X)^2 \to 0$$

which is often expressed as

$$X_n \overset{M}{\to} X.$$

For example, the sample mean $\overline{X}_n$ from a population with mean μ and variance σ^2 converges in mean square to μ because

$$E[(\overline{X}_n - \mu)^2] = \text{var}(\overline{X}_n) = \sigma^2 \big/ n \to 0 \text{ as } n \to \infty.$$

B.2.5 Convergence in Distribution

A sequence $\{X_n\}$, n = 1, 2, 3, . . . is said to converge to X if the distribution function F_n of X_n converges to the distribution function F of X at every continuity point of F. We express this as

(B.22)

$$X_n \xrightarrow{D} X$$

F is called the limit distribution of $\{X_n\}$.

For example, if $X \sim N(0,1)$, that is, it follows the standard normal distribution, we can write (B.22) as $X_n \xrightarrow{D} N(0,\ 1)$.

It may be noted that

$$\begin{aligned}
&\text{If } X_n \xrightarrow{D} X \text{ and } Y_n \xrightarrow{D} c,\ \text{ then } X_n + Y_n \xrightarrow{D} X + c \\
&\text{If } X_n \xrightarrow{D} X \text{ and } Y_n \xrightarrow{D} c,\ \text{ then } X_n Y_n \xrightarrow{D} cX \\
&\text{If } X_n \xrightarrow{D} X,\ \text{ then } g(X_n) \xrightarrow{D} g(X)
\end{aligned}$$

where c is a constant.

B.2.6 Relationships Among Three Modes of Convergence

There is a hierarchy among the various modes of convergence: **almost sure convergence** → **convergence in mean square** → **convergence in probability** → **convergence in distribution**, the arrow indicating which convergence dominates. This chain of convergence is not reversible: for example, convergence in distribution does not imply convergence in probability, and so on.[4] The various convergence theorems can be proved with the aid of Chebyshev's inequality and Khinchine's theorem. We will not prove this explicitly, for some of the proofs are involved and can be found in the references.

The concepts of convergence help us in distinguishing between the WLLN and the SLLN: the WLLN converges to μ in probability and the SLLN says that this is also true almost surely.

B.3 The Order of Magnitude of a Sequence

In studying the rate of convergence of sequences of variables, it is useful to know the concept of the **order of magnitude** of such sequences. We first consider the case of a nonstochastic sequence of real numbers a_n, as in the following examples.

© 2019 by DNG Econometrics, Inc

Consider the sequence $a_n = 6 + n - 4n^2$. As n becomes increasingly larger, the terms 6 and n become small (in absolute value) compared with $-4n^2$. We call the last term the *leading term* of the sequence, and it determines the order of magnitude of the sequence.

In general, we say that the sequence a_n is *at most* of order n^k, written as $O(n^k)$, or big O n^k, if the sequence $n^{-k}a_n$ is bounded. For the above sequence, if we take $k = 2$, we obtain

$$n^{-2}(6 + n + 4n^2) = \frac{6}{n^2} + \frac{1}{n} - 4$$

Therefore, as $n \to \infty$ this sequence converges to −4, so that it is bounded. Hence, this sequence is $O(n^2)$. As you can see, it is the leading term that determines the order of magnitude.

A related concept is *of smaller order than* n^k, written as $o(n^k)$, or small o (n^k), which means that the sequence $n^{-k}a_n$ converges to zero. In our example, $a_n = o(n^k)$ because

$$n^{-3}(6 + n - 4n^2) = \frac{6}{n^3} + \frac{1}{n^2} - \frac{4}{n}$$

As $n \to \infty$ this sequence tends to zero. This is true even if $k = 2.5$ or $k = -8$.

In the special case of $k = 0$, $a_n = O(n^0) = O(1)$, that is, the sequence is bounded, and in the case of small o, $a_n = o(n^0) = o(1)$, the sequence has a zero limit.

The algebra of sequences is similar to ordinary algebra. Thus, if $a_n = 4n$ and $b_n = 2 + n^{-1}$, then $a_n + b_n = 4n + 2 + n^{-1}$ and $a_nb_n = 8n + 4$.

In general, if $a_n = O(n^h)$ and $b_n = O(n^k)$, then $a_nb_n = O(n^{h+k})$ and

$$a_n + b_n = O(n^l)$$

where $l = \max(h,k)$.

B.4 The Order of Magnitude of a Stochastic Sequence

The ideas of order of magnitude for a nonstochastic sequence can be carried over to stochastic sequences.

Let X_n, $n = 1, 2, \ldots, N$ denote a sequence of real-valued random variables. Then X_n is said to be bounded in probability if for every ε &gt; 0, there exists a positive constant *c* and a positive integer *N* such that

(B.23)
$\Pr[\,|X_n|\;\;\&gt;; c] \leq \varepsilon$

for all $n \geq N$ and where ε is an arbitrarily small positive number.

In words, X_n is bounded in probability if for any arbitrarily small positive ε we can always find a positive constant *c* such that the probability of the absolute value of X_n being larger than *c* is less than ε.

Equation (B.23) can be equivalently expressed as

(B.24)
$\Pr[\,|X_n|\;\; \leq c] \&gt;; 1 - \varepsilon$

for all $n \geq N$

Notice the following statements:

1. $X_n = O_p(1)$ means that X_p is bounded in probability.
2. $X_n = O_p(1)$ means
   $X_n \xrightarrow{p} 0,$
   that is, $\Pr[\,|X_n|\;\;\&gt;; \varepsilon] \to 0$ as $n \to \infty$.

Notes

[1] See Amemiya, T. (1994). *Introduction to statistics and econometrics* (chap. 6). Cambridge, MA: Harvard University Press; Goldberger, A. S. (1991). *A course in econometrics* (chap. 9). Cambridge, MA: Harvard University Press.

[2] The inequality is also expressed as$\Pr(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2};\;\; k \&gt;; 0.$

[3] See Hogg, R. V., Mckean, J., &amp; Craig, A. T. (2014). *Introduction to mathematical statistics* (7th ed.). Noida, India: Pearson Education.

[4] For proof, see Ramanathan, R. (1993). *Statistical methods in econometrics* (chap. 7). New York, NY: Academic Press.

Appendix C: Small- and Large-Sample Properties of Estimators

Appendix C: Small- and Large-Sample Properties of Estimators

An estimator can be judged by several properties. These properties fall into two categories: (1) small sample and (2) large sample.

C.1 Small-Sample Properties of Estimators

C.1.1 Unbiasedness

An estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if the expected value of $\hat{\theta}$ is equal to θ, that is,

(C.1)

$$E(\hat{\theta}) = \theta$$

If this equality does not hold, then the estimator is said to be biased, which can be expressed as

(C.2)

$$\text{bias}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

It is important to note that unbiasedness is a property of repeated sampling, not of any given sample: Keeping the sample size fixed, we draw several samples from a given population, each time obtaining an estimate of the unknown parameter. For the estimator to be unbiased, the average value of these estimates must be equal to the true value.

C.1.2 Minimum Variance

$\hat{\theta}_1$ is said to be a minimum variance estimator of θ if its variance is smaller than or at most equal to the variance of$\hat{\theta}_2$, which is another estimator of *θ.* A minimum variance estimator is not necessarily unbiased.

C.1.3 Best Unbiased or Efficient Estimator

If $\hat{\theta}_1$ and $\hat{\theta}_2$ are two *unbiased* estimators of θ, and the variance of $\hat{\theta}_1$ is smaller than the variance of$\hat{\theta}_2$, then

© 2019 by DNG Econometrics, Inc

$\hat{\theta}_1$ is a **minimum variance unbiased**, or **best unbiased**, or **efficient**, estimator.

C.1.4 Linearity

An estimator $\hat{\theta}$ is said to be a linear estimator of θ if it is a linear function of the sample observations. For example, the sample mean defined as

$$\overline{X} = \frac{1}{n}\sum_{1}^{n} X_i = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

is a linear estimator because it is a linear function of the *X* values.

C.1.5 Best Linear Unbiased Estimator

If $\hat{\theta}$ is linear, is unbiased, and has minimum variance in the class of all linear unbiased estimators of θ, then it is called a best linear unbiased estimator, or BLUE for short.

C.1.6 Minimum Mean Square Error Estimator

The minimum mean square error (MSE) of an estimator $\hat{\theta}$ is defined as

(C.3)

$$\text{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2$$

whereas the variance of $\hat{\theta}$ is defined as

(C.4)

$$\text{var}(\hat{\theta}) = E[\hat{\theta} - E(\hat{\theta})]^2$$

The difference between the two is that the variance measures the dispersion of the distribution of an estimator around its mean or expected value, whereas MSE measures it around the true value of the parameter.

Simple algebraic manipulation will show that

(C.5)

$$
\begin{aligned}
MSE\left(\hat{\theta}\right) &= E\left[\hat{\theta} - E\left(\hat{\theta}\right)\right]^2 + \left[E\left(\hat{\theta}\right) - \theta\right]^2 \\
&= \operatorname{var}\left(\hat{\theta}\right) + \left[bias\left(\hat{\theta}\right)\right]^2
\end{aligned}
$$

Of course, if the bias is zero, MSE and variance of an estimator are the same. The MSE thus takes into account both bias and variance. Thus, MSE provides a trade-off between the variance and the bias of an estimator. In practice, the MSE criterion is used if the best unbiased criterion is incapable of producing estimators with smaller variances. In discussing multicollinearity in Chapter 7, we stated that estimators based on ridge regression are biased but they tend to have a smaller MSE than the OLS estimators.

C.1.7 Efficient Estimator

An estimator

$\hat{\theta}_1$

is an efficient estimator of θ if

(C.6)

$$E(\hat{\theta}_1) = \theta \text{ and } \operatorname{var}(\hat{\theta}_1) \leq \operatorname{var}(\hat{\theta}_2)$$

where $\hat{\theta}_2$ is any other unbiased estimator of θ.

An efficient estimator defined this way is also known as a *minimum variance unbiased estimator* (MVUE) or *best unbiased estimator.*

How does one find such an estimator? Here, we can get some guidance from the well-known Cramer–Rao (CR) theorem, which provides *a sufficient but not necessary condition* for an unbiased estimator to be efficient.[1]

To understand the idea behind the CR theorem, consider a random variable X with density function $f(X,\theta)$, where θ is the unknown parameter; for simplicity, we assume that there is only one unknown parameter, but this can be generalized for a density function with several unknown parameters. Let $X_1, X_2, \ldots, X_n$ denote a random sample drawn from this density function, n being the sample size.

Let $L(X_1, X_2, \ldots, X_n \mid \theta)$be the likelihood function of the sample, and further assume that it is twice differentiable and it satisfies certain **regularity conditions**, such as differentiation under the integral sign,

© 2019 by DNG Econometrics, Inc

the limit of integration is independent of the true θ, and the order of integration and differentiation can be interchanged. Let $\hat{\theta}$ be an unbiased estimator of θ. Then the variance of $\hat{\theta}$ must satisfy the inequality

(C.7)

$$\operatorname{var}(\hat{\theta}) \geq -\frac{1}{E\left(\frac{\partial^2 l}{\partial \theta^2}\right)}$$

where l = ln L, that is, the (natural) log of the likelihood function, L. The right-hand side of this equation is known as the Cramer–Rao lower bound (CRLB).

The quantity $\boldsymbol{I} = E\left(\frac{\partial^2 \ln l}{\partial \theta^2}\right)$ is called **Fisher's information matrix**[2]—the amount of information that a sample provides about the value of an unknown parameter(s). Thus, the greater the variance, the less the information.

The basis of the CR inequality is Fisher's information matrix, although since 1948 it has been called the CR inequality because Cramer and Rao proved this inequality independently. An estimator that achieves the CRLB is called an *efficient estimator*.

The CR inequality can be generalized to account for more than one unknown parameter, say, $\theta_1, \theta_2, \ldots, \theta_p$, in which case the information matrix can be expressed as

(C.8)

$$\boldsymbol{I}_{jk} = E\left\{\left(\frac{\partial \ln l}{\partial \theta_j}\right)\left(\frac{\partial \ln l}{\partial \theta_k}\right)\right\}'$$

Commenting on the CR inequality, Theil writes as follows:

> As soon as we have found an unbiased estimator whose variance is equal to minus the reciprocal of the expected second-order derivative of the log-likelihood of the sample, we know that we cannot find an unbiased estimator with a smaller variance. Hence, no further improvements are possible if we confine ourselves to unbiased estimators and if minimum sample variance is our goal.[3]

However, an efficient estimator in this sense may not exist. This is because the CRLB is only a sufficient condition and not a necessary one because in some situations the lower bound on the variance cannot be attained by any unbiased estimator. Besides, one or more regularity conditions underlying the CR theorem may not be satisfied. However, in some cases, we can actually find the variance of an estimator that satisfies the CR inequality, as the following example shows.

Consider the sample mean $\overline{X}$ from a normal population with mean μ. We want to prove that $\overline{X}$ is an MVUE of μ.

Proof: Since

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \quad -\infty \text{ \&lt; } X \text{ \&lt; } \infty$$

it follows that

$$\ln f(X) \;=\; -\ln\sigma\sqrt{2\pi} - \frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2$$

Now

$$\frac{\partial \ln f(X)}{\partial \mu} \;=\; \frac{1}{\sigma}\left(\frac{X-\mu}{\sigma}\right)$$

Hence,

$$E\left[\left(\frac{\partial \ln f(X)}{\partial \mu}\right)^2\right] \;=\; \frac{1}{\sigma^2}\cdot E\left[\left(\frac{X-\mu}{\sigma}\right)^2\right] \;=\; \frac{1}{\sigma^2}\cdot 1 = \frac{1}{\sigma^2}$$

Therefore,

(C.9)

$$\frac{1}{n\cdot E\left[\left(\frac{\partial \ln f(X)}{\partial \mu}\right)^2\right]} \;=\; \frac{1}{n\cdot\frac{1}{\sigma^2}} \;=\; \frac{\sigma^2}{n}$$

Since $\overline{X}$ is unbiased and $\operatorname{var}(\overline{X}) \;=\; \frac{\sigma^2}{n},$ according to the CR theorem, $\overline{X}$ is an MVUE of μ.

We have shown earlier that in the classical normal linear regression, the ML estimators of the regression parameters attain the CRLB and hence they are most efficient.

C.1.8 Sufficient Estimator(s)

Consider a random variable X with the density function $f(X,\theta)$, where θ is a single parameter of this density function; for simplicity of exposition, we assume that there is just a single parameter, although we can extend it to multiple parameters. Suppose from this density function, we draw a random sample of n observations, X_1, $X_2, \ldots, X_n$ and compute a sample statistic, say $\hat{\theta}$. We say that $\hat{\theta}$ is a *sufficient* statistic for θ if it *encapsulates* all the information about θ, that is, it condenses the sample data in such a way that no information about θ is lost. If such a statistic exists, there is no need to look for examining the entire sample or another statistic based on this sample.

Formally, let X ={$x1$, $x2$, . . . , xn} be a random sample from a population with density function Then the statistic or estimator $\hat{\theta}$ is a sufficient estimator of θ if and only if the joint density or probability distribution of the random sample can be factored so that

(C.10)

$$L(\theta; \boldsymbol{X}) = f(X_1, X_2, \ \ldots, X_n; \theta) = h(\theta, \hat{\theta}) \cdot g(\boldsymbol{X})$$

where $g(\boldsymbol{X})$ stands for $g(\boldsymbol{X}1, \boldsymbol{X}2, \ldots, \boldsymbol{X}n)$ and where $h(\theta, \hat{\theta})$ depends only on $\hat{\theta}$ and θ and $g(\boldsymbol{X})$ does not depend on θ.

Equation (C.10) is known as the **Fisher–Neyman factorization theorem**. This theorem states that a statistic or estimator $\hat{\theta}$ is sufficient for θ if and only if the joint density of the sample can be factored into two components—one depends only on the estimator and the true parameter and the other is independent of the parameter.

Example: Consider a sample of *n* observations from a normal population with mean μ and *known* variance σ^2 It can be shown that $\bar{X}$, the sample mean, is a sufficient estimator of μ.

Since

(C.11)

$$L(X_1, X_2, \ \ldots, X_n) \ = \ \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{\left(-\frac{1}{2}\right)\Sigma\left(\frac{X_i - \mu}{\sigma}\right)^2\right\}$$

the log-likelihood function becomes

(C.12)

$$l(X_1, X_2, \ \ldots, X_n) \ = \ n\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\Sigma\left(\frac{X_i - \mu}{\sigma}\right)^2$$

With simple algebraic manipulation, it can be shown that

(C.13)

$$\Sigma(X_i - \mu)^2 \ = \ \Sigma(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$$

© 2019 by DNG Econometrics, Inc

As a result, we can express Equation (C.12) as

(C.14)

$$l(X_1, X_2, \ \dots , X_n) = \ln\left(\frac{\sqrt{n}}{\sigma\sqrt{2\pi}}\right) - \frac{1}{2}\left(\frac{\bar{X} - \mu}{\sigma\sqrt{n}}\right)^2$$

$$+\left\{\ln\frac{1}{\sqrt{n}}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{n-1} - \frac{1}{2}\Sigma\left(\frac{X_i - \bar{X}}{\sigma}\right)^2\right\}$$

From Equation (C.14), we see that the first factor on the right-hand side involves $\bar{X}$ and the population mean μ, and the second factor does not involve *μ.* By the Fisher–Neyman factorization theorem, we can therefore say that $\bar{X}$ is a sufficient estimator of μ of a normal population with the known variance σ^2.

C.1.8.1 Properties of Sufficient Estimators

It is important to know some of the properties of sufficient statistics:[4]

1. If $\hat{\theta}$ is sufficient for θ, then the likelihood function for θ based on the distribution of $\hat{\theta}$ is proportional to *L*(θ;*X*), where *L*(θ;*X*), is given by (C.10).

   This result makes sense, for $\hat{\theta}$ carries all the sample information about θ. In other words, we get the same information about θ from $\hat{\theta}$ as we would get from the entire sample of ***y***.

2. If $\hat{\theta}$ is sufficient for θ, the conditional distribution of outcomes ***y***, given the observed value of $\hat{\theta}$, does not depend on θ. A sufficient statistic is often defined this way.
3. Every single-valued function of the sufficient statistic is also a sufficient statistic of the true value of the estimator. Stated differently, a one-to-one transformation of sufficient statistics produces another set of sufficient statistics. For example, if $X_1, X_2, \dots , X_n$ are independent Bernoulli random variables with the parameter θ, then

   $$\hat{\theta} = \frac{\Sigma_1^n X_i}{n}$$

   is a sufficient estimator of θ and $X = X_1 + X_2 + \dots + X_n$ is also a sufficient estimator of the binomial mean μ = *nθ.*

4. As Kalbfleisch notes, the ML estimator $\hat{\theta}$ is part of any set of sufficient statistics $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$, in that its value can be computed from just the $\hat{\theta}_i s$. This is because the $\hat{\theta}_i s$ determine $L(\theta,y)$ up to a proportionality constant, and $\hat{\theta}$ does not depend on this constant.

C.1.9 Uniformly Minimum Variance Unbiased Estimator

An unbiased estimator that has minimum variance in the entire class of unbiased estimators, whether linear or nonlinear, is called a uniformly minimum variance unbiased estimator (UMVUE). More technically, an unbiased estimator that has variance equal to CRLB must have minimum variance among all unbiased estimators. Sufficiency is a powerful property in finding UMVUE estimators.

C.2 Large-Sample Properties of Estimators

An estimator may not satisfy one or more of the desirable statistical properties in small samples, but as the sample size increases indefinitely, the estimator may possess several desirable statistical properties. These properties are known as **large-sample**, or **asymptotic**, **properties**, such as the following:

C.2.1 Asymptotic Unbiasedness

An estimator $\hat{\theta}$ is said to be an asymptotically unbiased estimator of θ if

(C.15)

$$\lim_{n \to \infty} E(\hat{\theta}_n) = \theta$$

where the subscript n on the estimator means that the estimator is based on a sample size n and where “lim” means limit and $n \to \infty$ means that n increases indefinitely. In words, $\hat{\theta}$ is an asymptotically unbiased

© 2019 by DNG Econometrics, Inc

estimator of θ if its expected, or mean, value approaches the true value as the sample size gets larger and larger.

As an example, consider the following measure of the sample variance of a random variable *X* with mean of μ and variance σ^2:

(C.16)

$$S^2 = \frac{\Sigma (X_i - \bar{X})^2}{n}$$

Algebraic manipulation will show that

$$\begin{aligned} \sum \left(X_i - \bar{X}^{2}\right)^2 &= \sum \left[(X_i - \mu) - \left(\bar{X} - \mu\right)\right]^2 \\ &= \sum (X_i - \mu)^2 - n\left(\bar{X} - \mu\right)^2 \end{aligned}$$

Therefore,

(C.17)

$$\begin{aligned} S^2 &= \frac{\Sigma (X_i - \mu)^2 - n(\bar{X} - \mu)^2}{n} \\ &= \frac{E\Sigma (X_i - \mu)^2 - nE(\bar{X} - \mu)^2}{n} \\ E(S^2) &= \sigma^2 - \frac{\sigma^2}{n} \\ &= \sigma^2\left(1 - \frac{1}{n}\right) \end{aligned}$$

which shows that the sample variance as defined in (C.16) is a biased estimator of σ^2*Note:* $E(X_i - \mu)^2 = \sigma^2$ and $E(\bar{X} - \mu)^2 = \sigma^2 \big/ n,$ which is the variance of the sample mean.

However,

(C.18)

$$\underset{n \to \infty}{E(S^2)} = \sigma^2$$

Obviously, S^2 is biased, but as *n* increases indefinitely, $E(S^2)$ approaches the true σ^2, hence it is asymptotically unbiased. We have showed earlier that the ML estimator of the error variance of the normal LRM is biased, but the bias gets smaller and smaller as the sample size increases indefinitely.

It may be noted that if we define the sample variance as

(C.19)

$$s^2 = \frac{\Sigma(X_i - \overline{X})^2}{n - 1}$$

then it can be shown that this estimator is unbiased, as $E(s^2) = \sigma^2$ regardless of the sample size. Note that we have used s^2 to distinguish it from the S^2 defined in (C.16); the difference between the two lies in that (C.19) takes into account the degrees of freedom.

$$(n - 1)s^2 = \sum \left[X_i - \mu - \left(\overline{X} - \mu\right)\right]^2$$

$$= \sum \left[\left(X_i - \mu\right)^2 - s\left(X_i - \mu\right)\left(\overline{X} - \mu\right) + \left(\overline{X} - \mu\right)^2\right]$$

$$= \sum \left(X_i - \mu\right)^2 - n\left(\overline{X} - \mu\right)^2$$

since

$$2\Sigma(X_i - \mu)(\overline{X} - \mu) = 2n(\overline{X} - \mu)^2$$

Now,

$$(n - 1)Es^2 = E\sum \left(X_i - \mu\right)^2 - nE\left(\overline{X} - \mu\right)^2$$

$$= n\sigma^2 - n(\sigma^2 / n)$$

$$= n\sigma^2 - \sigma^2$$

$$= \sigma^2(n - 1)$$

Therefore,

(C.20)

$$Es^2 = \sigma^2$$

In deriving this result, we make use of an important relationship that for any random variable, say, *V*,

(C.21)

© 2019 by DNG Econometrics, Inc

$$E(V^2) = \mathrm{var}(V) + (E[V])^2$$

C.2.2 Consistency

The property of consistency is concerned with the asymptotic (i.e., as $n \to \infty$) accuracy of an estimator, that is, whether it converges to the parameter that it is estimating.

$\hat{\theta}$ is said to be a consistent estimator of the population parameter θ if it approaches the true value θ as the sample size gets larger and larger. More formally, an estimator $\hat{\theta}$ is a consistent estimator of θ if the probability that the absolute value of the difference between the two is less than δ (an arbitrarily small positive quantity) and approaches unity. Symbolically,

(C.22)

$$\lim_{n \to \infty} \Pr\left\{ \left| \hat{\theta} - \theta \right| \text{\&lt;; } \delta \right\} = 1 \ \delta \text{ \&gt;; } 0$$

where Pr stands for probability. This is often expressed as

(C.23)

$$p\lim_{n \to \infty} \hat{\theta} = \theta$$

where *p*lim means probability limit.

It is important to note that the properties of unbiasedness and consistency are conceptually quite different. The property of unbiasedness can hold for any sample size, whereas consistency is strictly a large-sample property.

A *sufficient condition* for consistency is that the bias and variance both tend to zero as the sample size increases indefinitely. Technically,

(C.24)

$$\lim_{n \to \infty} E(\hat{\theta}) = \theta \text{ and } \lim_{n \to \infty} \mathrm{var}(\hat{\theta}_n) = 0$$

Alternatively, a sufficient condition for consistency is that the

$\mathrm{MSE}(\hat{\theta})$

tends to zero as *n* increases indefinitely.

© 2019 by DNG Econometrics, Inc

Example: Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. It is easy to show that the sample mean

$\bar{X}$

is a consistent estimator of μ. Since

$E(\bar{X}) = \mu$

regardless of the sample size, it is unbiased and

$\mathrm{var}(\bar{X}) = \sigma^2 / n$

regardless of the sample size. Furthermore, as n increases indefinitely,

$\mathrm{var}(\bar{X})$

tends toward zero. Hence, the sample mean is a consistent estimator of the population mean. In short, $p\lim(\bar{X}) = \mu$.

C.2.2.1 Probability Limit (plim)

In establishing the consistency property of estimators, the following properties of the probability limit ($p$lim) are noteworthy.

1. *Invariance* (Slutsky property). If $\hat{\theta}$ is a consistent estimator of θ and if

   $h(\hat{\theta})$

   is any continuous function of $\hat{\theta}$, then

   (C.25)

   $$\underset{n \to \infty}{p\lim}\, h(\hat{\theta}) = h(\theta)$$

   What this means is that if $\hat{\theta}$ is a consistent estimator of θ, then

   $1 / \hat{\theta}$

   is also a consistent estimator of 1/θ and that

   $\log(\hat{\theta})$

   is a consistent estimator of log(θ)

   The invariance property does not hold true of the expectations operator E. Thus, if $\hat{\theta}$ is an unbiased estimator of θ, it is not the case that

   $1 / \hat{\theta}$

   is an unbiased estimator of 1/θ, that is

© 2019 by DNG Econometrics, Inc

(C.26)

$$E\left(\frac{1}{\hat{\theta}}\right) \neq \frac{1}{E(\hat{\theta})} \neq \frac{1}{\theta}$$

2. If b is a constant, then

   (C.27)

   $$\underset{n \to \infty}{\text{plim}} = b$$

   That is, the probability limit of a constant is the same constant.

3. If $\hat{\theta}_1$ *and* $\hat{\theta}_2$ are consistent estimators, then

(C.28)

$$plim(\hat{\theta}_1 + \hat{\theta}_2) = plim\hat{\theta}_1 + plim\hat{\theta}_2$$

$$plim(\hat{\theta}_1 - \hat{\theta}_2) = plim\hat{\theta}_1 - plim\hat{\theta}_2$$

$$plim(\hat{\theta}_1\hat{\theta}_2) = plim\hat{\theta}_1 plim\hat{\theta}_2$$

$$plim\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right) = \frac{plim\hat{\theta}_1}{plim\hat{\theta}_2}; \quad \text{for } plim\hat{\theta}_2 \neq 0$$

Again, note that the last two properties do not hold true of the expectations operator *E*. Thus,

(C.29)

$$E\left(\frac{\hat{\theta}_1}{\hat{\theta}_2}\right) \neq \frac{E(\hat{\theta}_1)}{E(\hat{\theta}_2)}$$

$$E(\hat{\theta}_1\hat{\theta}_2) \neq E(\hat{\theta}_1)E(\hat{\theta}_2)$$

If, however, $\hat{\theta}_1$ *and* $\hat{\theta}_2$ are independently distributed,

(C.30)

$$E(\hat{\theta}_1\hat{\theta}_2) = E(\hat{\theta}_1)E(\hat{\theta}_2)$$

Given the classical setup, we have shown that the OLS estimators are unbiased. We can also show that these estimators are consistent. To illustrate, consider the following simplest linear regression:

$Y_i = B_1 + B_2X_i + u_i$

Under classical assumptions, it is easy to show that

$$b_1 = \overline{Y} - b_2\overline{X}$$
$$b_2 = \frac{\Sigma y_i x_i}{\Sigma x_i^2}$$

where

$$y_i = (Y_i - \overline{Y});\ \ x_i = (X_i - \overline{X})$$
$$\text{var}(b_2) = \frac{\sigma^2}{\Sigma x_i^2}$$

To prove that $b2$ is a consistent estimator of $B2$, we need to show that the variance of $b2$ tends to zero as n, the number of sample observations, increases indefinitely. $\Sigma x_i^2 \,/\, n \neq 0$, because the variance of X is bounded. We proceed as follows:

$$\text{var}(b_2) = \frac{\sigma^2}{\Sigma x_i^2} = \frac{\sigma^2 \,/\, n}{\Sigma x_i^2 \,/\, n}$$

Dividing the numerator and the denominator by n, we do not change the equality.

Now

(C.31)

$$\lim_{x \to \infty} \text{var}(b_2) = \lim_{x \to \infty}\left(\frac{\sigma^2 \,/\, n}{\Sigma x_i^2 \,/\, n}\right) = \frac{\lim(\sigma^2 \,/\, n)}{\lim(\Sigma x_i^2 \,/\, n)} = 0$$

In establishing the preceding, we make use of the following properties: (1) The limit of a ratio quantity is the limit of the quantity in the numerator to the limit of the quantity in the denominator, (2) as n increases indefinitely σ^2 / n tends to zero because σ^2 is a finite number, and (3)
$\Sigma x_i^2 \,/\, n \neq 0$
because the variance of X has a finite limit because of the assumption of the classical linear regression.

We leave it to the reader to show that $b1$ is also a consistent estimator of $B1$. (Note thatvar$(b_1) = \frac{\Sigma X_i^2}{n\Sigma x_i^2}\sigma^2$.)

We can extend this analysis to multiple regression: $\boldsymbol{y} = \boldsymbol{XB} + \boldsymbol{u}$.

We have already shown that the estimator $\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ is an unbiased estimator of **$\boldsymbol{B}$ and that variance is $\boldsymbol{\sigma^2(X'X)}$**. To prove that $\boldsymbol{b}$ is a consistent estimator, we proceed as follows:

$$\begin{aligned}\text{var}(b) &= \sigma^2(X'X)^{-1}\\ &= \frac{\sigma^2}{n}\left(n^{-1}\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\end{aligned}$$

Taking the probability limit of this expression as $n \to \infty$, we obtain

(C.32)

$$\begin{aligned}\lim_{n\to\infty}\text{var}(\boldsymbol{b}) &= \lim_{n\to\infty}\left[\frac{\sigma^2}{n}(n^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1}\right]\\ &= \lim_{n\to\infty}\frac{\sigma^2}{n}\lim_{n\to\infty}\left(n^{-1}X'X\right)^{-1}\end{aligned}$$

The assumption that the elements of the matrix $\boldsymbol{X}$ are bounded implies that $n^{-1}\boldsymbol{X'X}$ and hence $(n^{-1}\boldsymbol{X'X})^{-1}$ is also bounded for all n. Therefore, $\lim\limits_{n\to\infty}(n^{-1}\boldsymbol{X}'\boldsymbol{X})^{-1}$ can be replaced by a matrix of finite constants since $\lim\limits_{n\to\infty}\frac{\sigma^2}{n} = 0$. This establishes the consistency of $\boldsymbol{b}$.

C.2.2.2 Consistency of S^2:[5]

If b is a consistent estimator of $\boldsymbol{B}$ and the matrix $\boldsymbol{X}$ is nonstochastic, we can prove that S^2 is a consistent estimator of the true variance, σ^2 To show this, we proceed as follows:

(C.33)

$$\begin{aligned}e &= y - Xb\\ &= (XB + u) - Xb\\ &= X(B - b) + u\\ (e - u) &= X(B - b)\end{aligned}$$

Now each element ($\boldsymbol{e} - \boldsymbol{u}$) converges in probability to zero. That is, e_t converges in probability to u_t. As a result, the limiting behavior of $S^2 = \Sigma e_t^2 / n - k$ is equal to the limiting behavior of

$\Sigma u_t^2 / n - k$,

which is equal to the limiting behavior of $\Sigma u_t^2 / n$. Since the variables u_t. are iid with $E(u_t^2) = \sigma^2$, then by Khinchine's theorem it follows that $p\lim(\Sigma u_t^2 / n) = \sigma^2$, which proves that S^2 is a consistent estimator of σ^2.

C.2.3 Asymptotic Efficiency

Let $\hat{\theta}$ be an estimator of θ. The variance of the asymptotic distribution of $\hat{\theta}$ is called the **asymptotic variance** of $\hat{\theta}$. If $\hat{\theta}$ is consistent and its asymptotic variance is smaller than the asymptotic variance of all other consistent estimators of θ, then $\hat{\theta}$ is called **asymptotically efficient**.

C.2.4. Asymptotic Normality

An estimator $\hat{\theta}$ is said to be asymptotically normally distributed if its sampling distribution approaches the normal distribution as the sample size n increases indefinitely. For example, statistical theory shows that if $X_1, X_2, \ldots, X_n$ are independently normally distributed with the same mean μ and the same variance σ^2, the sample mean $\bar{X}$ is also normally distributed with mean μ and variance σ^2 / n in small as well as large samples. But if values of X_i are independent with mean μ and variance σ^2 but are not necessarily from the normal distribution, then the sample mean $\bar{X}$ is *asymptotically* normally distributed with mean μ and variance σ^2 / n. This in essence is one version of the **central limit theorem (CLT)**, which is of fundamental importance in statistics.

Let $X_1, X_2, \ldots, X_n$ represent a random sample from any population with $E(X)=\mu$ and $\text{var}(x) = \sigma^2$consider the standardized sample mean

(C.34)

$$Z = \frac{\left(\bar{X} - \mu\right)}{\sigma / \sqrt{n}} = \frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma}$$

By the CLT, Z converges in distribution to $N(0, 1)$. Equivalently,

$\sqrt{n}(\overline{X} - \mu)$

converges in distribution to $N(0, \sigma^2)$.[6]

As Goldberger notes, “Approximating the cdf of the standardized sample mean Z_n by the $N(0, 1)$ cdf amounts to approximating the cdf of the sample mean

$\overline{X}_n$

by the $N(\mu, \sigma^2 \big/ n)$ cdf.”[7] As a result, we can write

(C.35)

$$\overline{X} \sim \text{asy}N(\mu, \sigma^2 \big/ n)$$

In other words, the *asymptotic distribution of* $\overline{X}_n$ is normal with mean μ and variance σ^2 / n and we can call μ and σ^2 / n as the *asymptotic expectation* and *asymptotic variance* of the estimator $\overline{X}_n$, where n is the sample size.

It is important to note the distinction between the *limiting* distribution of the sample mean, which is degenerate at μ (see our discussion about the consistency of an estimator), and the *asymptotic distribution* of the sample mean, which is $N(\mu, \sigma^2 / n)$. As Goldberger notes, the latter provides more useful information.

In case a density function involves more than one parameter, we can express asymptotic normality in the following form:

(C.36)

$$\hat{\boldsymbol{\theta}} \sim \text{asy}N[\theta, \boldsymbol{I}^{-1}(\theta)]$$

What (C.36) says is that the estimated vector of parameters is normally distributed with mean the population parameter vector θ and the variance given by the inverse of the *information matrix* $\boldsymbol{I}(\theta)$, which is defined as

(C.37)

$$\boldsymbol{I}(\theta) = E\left[\left(\frac{\partial l}{\partial \theta}\right)\left(\frac{\partial l}{\partial \theta}\right)'\right] = -E\left[\frac{\partial^2 l}{\partial \theta \partial \theta'}\right]$$

where l is the log-likelihood function.

If θ is a vector of k elements, $\frac{\partial l}{\partial \theta}$ is a column vector of k partial derivatives, that is,

(C.38)

$$\frac{\partial lf}{\partial \theta} = \begin{bmatrix} \partial lf / \partial \theta_1 \\ \partial lf / \partial \theta_2 \\ \vdots \\ \partial lf / \partial \theta_k \end{bmatrix}$$

Each element in this **score** or **gradient vector** is itself a function of θ and may be differentiated partially with respect to each element in θ. As an example,

(C.39)

$$\frac{\partial}{\partial \theta_1}(\partial l \Big/ \partial \theta) = \begin{bmatrix} \frac{\partial^2 l}{\partial \theta_1^2} \\ \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} \\ \vdots \\ \frac{\partial^2 l}{\partial \theta_1 \partial \theta_k} \end{bmatrix}$$

Proceeding in this manner, we obtain the following square symmetric matrix of second-order derivatives, which is known as the **Hessian matrix** that we encountered earlier.

(C.40)

$$\frac{\partial^2 l}{\partial \theta \partial \theta'} = \begin{bmatrix} \frac{\partial^2 l}{\partial^2 \theta_1^2} & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 l}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 l}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_2^2} & \cdots & \frac{\partial^2 l}{\partial \theta_2 \partial \theta_k} \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\partial^2 l}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 l}{\partial \theta_k \partial \theta_2} & \cdots & \frac{\partial^2 l}{\partial \theta_k^2} \end{bmatrix}$$

Notes

[1] The proof of the CR theorem is rather involved and can be found in Theil, H. (1971). *Principles of econometrics* (pp. 384–386). New York, NY: John Wiley.

[2] Sir Ronald Fisher, a British statistician, developed this matrix in 1922.

[3] See Theil, H. (1971). *Principles of econometrics* (pp. 386–387). New York, NY: John Wiley.

[4] For details, see Kalbfleisch, J. G. (1985). *Probability and statistical inference, vol. 2: Statistical inference* (2nd ed., pp. 285–289). New York, NY: Springer-Verlag.

[5] The following discussion is based on Stewart, J., &amp; Gill, L. (1998). *Econometrics* (2nd ed., p. 114). Upper Saddle River, NJ: Prentice Hall.

[6] For proof, see DeGroot, M. H. (1975). *Probability and statistics.* Reading, MA: Addison-Wesley. The proof involves moment-generating or characteristic functions. It may be added that there are several versions of the CLT.

[7] Goldberger, A. S. (1991). *A course in econometrics* (p. 99). Cambridge, MA: Harvard University Press.

Appendix D: Some Important Probability Distributions1

Appendix D: Some Important Probability Distributions[1]

Estimation and hypothesis testing are two important branches of *statistical inference*. We have discussed estimation of the linear regression model using ordinary least squares and maximum likelihood methods of estimation. Hypothesis testing requires developing appropriate **test statistics** and their probability distributions that will aid us in testing hypotheses and establishing confidence intervals for the parameters of interest. We now discuss several probability distributions and their properties that will help us in developing appropriate test statistics.[2]

D.1 The Normal Distribution and the *Z* Test

Let $X_i \sim N(\mu, \sigma^2)$, that is, a normally distributed random variable with mean μ and variance σ^2 Its probability density function is

(D.1)

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}}\exp^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2} \text{ where}\sigma \text{ &gt; } 0$$

Now consider the following variable:

(D.2)

$$Z_i = \frac{X_i - \mu}{\sigma}$$

Z_i is a **standardized normal variable**, which has zero mean and unit variance, that is,

(D.3)

$$Z_i \sim N(0, 1)$$

D.1.1 Properties of the Normal Distribution

1. It is symmetric around its mean value.
2. Approximately 68% of the area under the normal curve lies between the values of μ ± σ, about 95% of the area lies between μ ± 2σ, and about 99.7% of the area lies between μ ±

3σ, (see Figure D.1).

3. It depends on only two parameters, mean μ and variance σ^2.
4. If $X_1 \sim N(\mu_1, \sigma_1^2)$ *and* $X_2 \sim N(\mu_2, \sigma_2^2)$, and they are independent, then

(D.4)

$$Y = (aX_1 + bX_2) \sim N[(a\mu_1 + b\mu_2), (a^2\sigma_1^2 + b^2\sigma_2^2)]$$

In words, *a linear combination of normally distributed variables is also normally distributed*. This result can be generalized to linear combinations of more than two independently distributed normal variables.

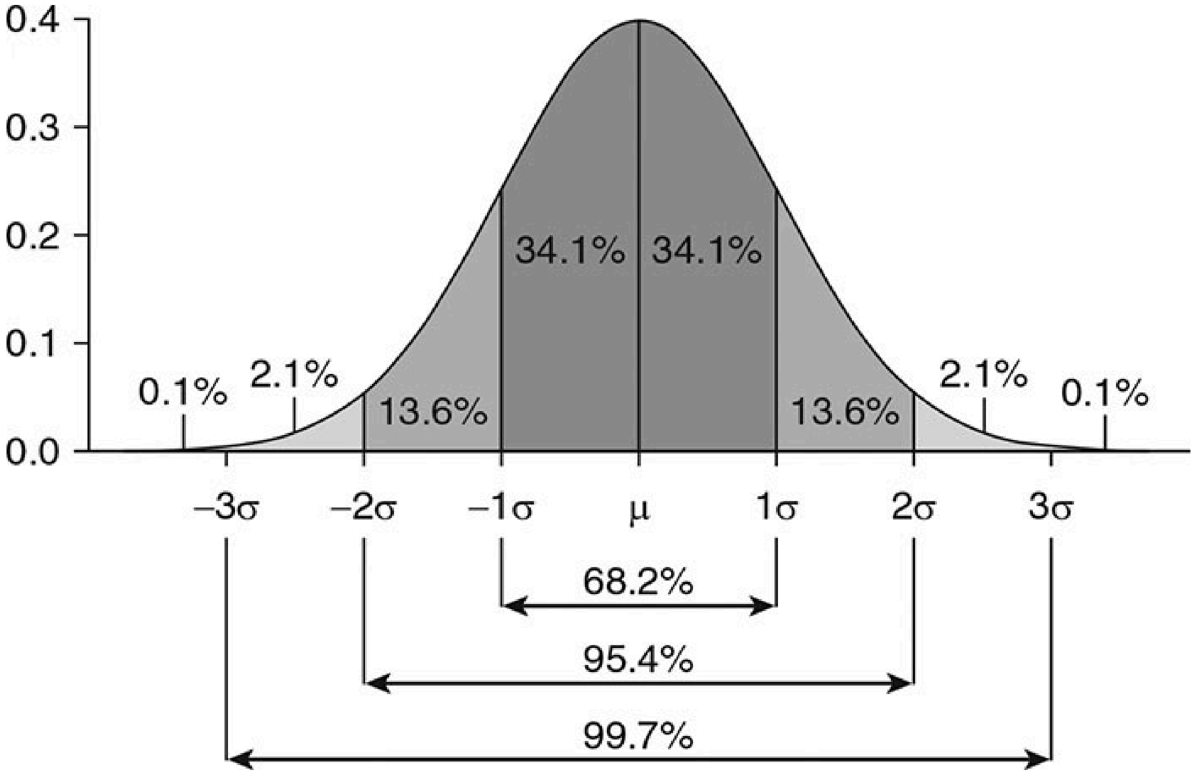
5. The third and fourth moments of the normal distribution around the mean value are as follows:

$$\textit{Thirdmoment}: E(X - \mu)^3 = 0$$

$$\textit{Fourthmoment}: E(X - \mu)^4 = 3\sigma^4$$

Note: All odd-powered moments about the mean value of a normally distributed variable are zero.

Figure D.1 The Normal Distribution



6. As a result, for a normally distributed variable, the **skewness coefficient** (*S*) is zero and the **kurtosis coefficient** (*K*) is 3. Skewness is a measure of asymmetry of a probability distribution, and

© 2019 by DNG Econometrics, Inc

kurtosis is a measure of how tall or flat the probability distribution is.

7. A simple test of normality is to find out whether the skewness coefficient and kurtosis measures are 0 and 3, respectively. This is the basis of the **Jarque–Bera (JB) test of normality**, which is defined as follows:

(D.5)

$$\mathrm{JB} = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right] \sim X_2^2$$

Under the null hypothesis of normality, the JB statistic is distributed as a chi-square statistic with 2 *df*. Notice that the JB statistic is a test of the *joint hypothesis* that the skewness coefficient is zero and the kurtosis coefficient is 3. That is the reason for the 2 *df*.

8. The mean and variance of a normally distributed random variable are independent, that is, one is not a function of the other.

9. If X and Y are jointly normally distributed, then they are independent, if and only if the covariance between them is zero.

The central limit theorem (CLT): Let $X_1, X_2, \ldots, X_n$ denote n independent random variables, all of which have the same probability distribution function with mean = μ and variance = σ^2 Let $\bar{X} = \Sigma X_i \big/ n$ be the sample mean. Then as $n \to \infty$,

(D.6)

$$\bar{X} \sim N(\mu, \sigma^2 \big/ n)$$

That is,

$\bar{X}$

approaches the normal distribution with mean μ and variance σ^2 / n. This result holds true regardless of the form of the probability distribution function. As a result, it follows that

(D.7)

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0, 1)$$

That is, Z is a standardized normal variable.

D.2 The Gamma Distribution

A random variable *X* has a gamma distribution if its probability distribution function is as follows:

(D.8)

$$f(X) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}X^{\alpha - 1}\exp(- X \Big/ \beta) \text{ for } X \text{ \&gt; } 0$$
$$= 0 \text{ elsewhere}$$

where α &gt; 0 and β &gt;0 (see Figure D.2).

The gamma function, Γ(α), is defined as

(D.9)

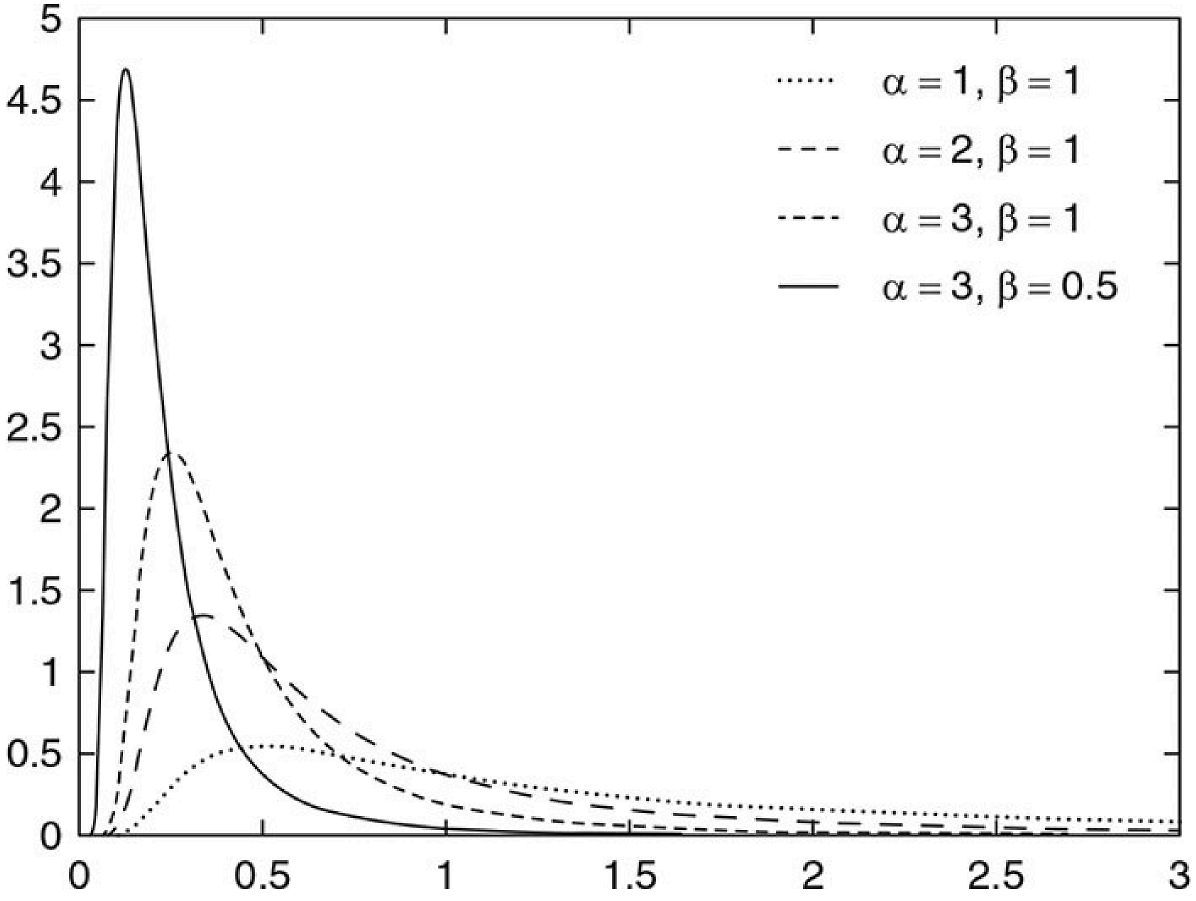
$$\Gamma(\alpha) = \int_0^{\infty} Y^{\alpha - 1}e^{-Y}dy \text{ for } \alpha \text{ \&gt; } 0$$

Using calculus techniques (integration by parts), it can be shown that the gamma function satisfies the recursive formula

(D.10)
$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1) \text{ for} \alpha \text{ \&gt; } 1$$

© 2019 by DNG Econometrics, Inc

Figure D.2 The Gamma Distribution



Note that

(D.11)

$$\Gamma(1) = \int_0^\infty e^{-Y} dy = 1$$

Then, by repeated application of the recursive formula it follows that

(D.12)

$$\Gamma(\alpha) = (\alpha - 1)!$$

where α is a positive integer.

An important special case is

(D.13)

$$\Gamma(\frac{1}{2}) = \sqrt{\pi}$$

© 2019 by DNG Econometrics, Inc

Special cases of the gamma distribution are the **chi-square**, **Erlang**, and **exponential distributions**.

D.3 The Chi-Square (χ^2) Distribution and the χ^2 Test

Let $Z_1, Z_2, \ldots, Z_k$ be independent standardized normal variables [ie., $Z_i \sim N(0,1)$].

Then, the quantity

(D.14)

$$Z = \Sigma Z_i^2 \sim \chi_k^2$$

is said to possess the χ^2 distribution with *k df*, where the *df* means the number of independent quantities in the previous sum.

A test statistic based on the chi-square distribution is called a **chi-square test**.

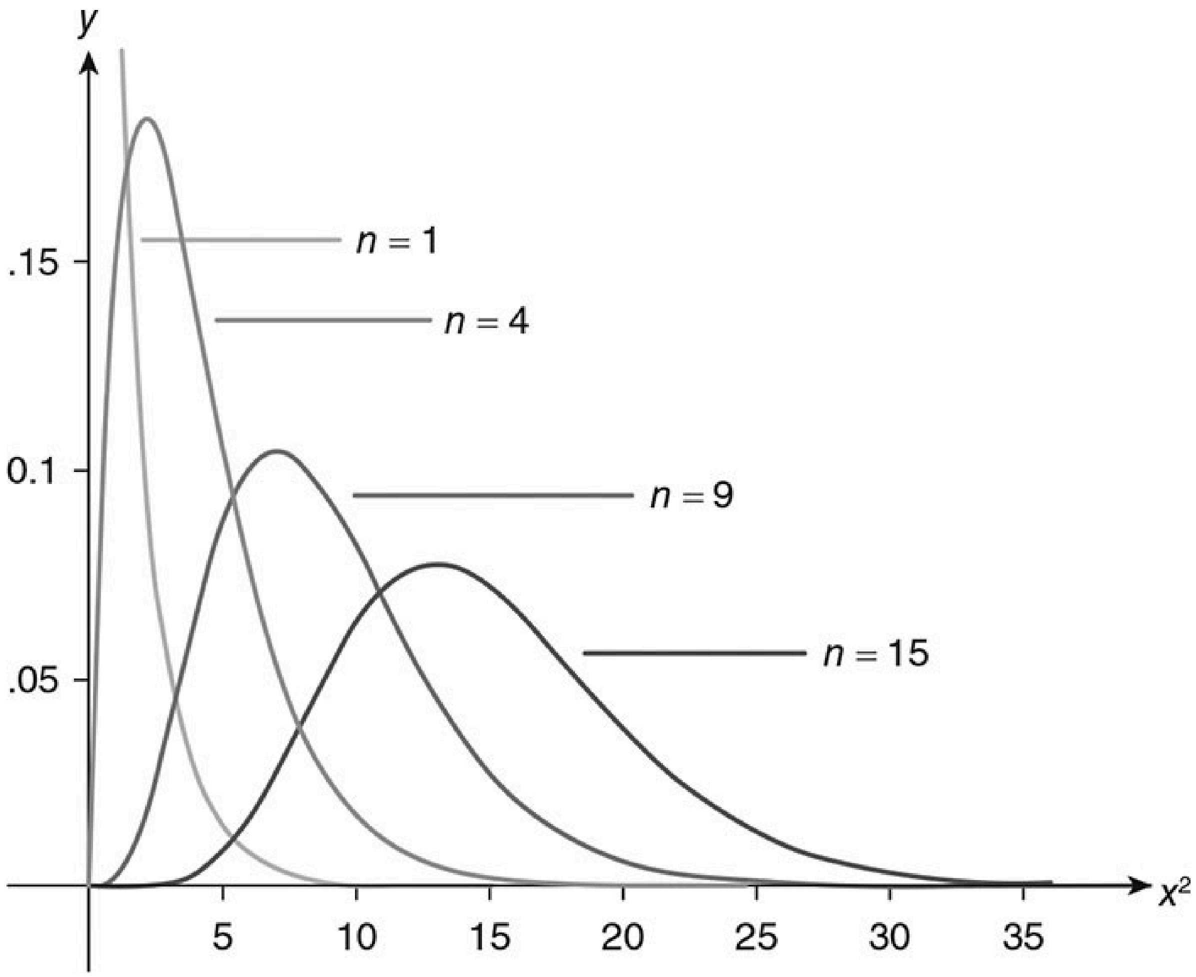
The probability distribution function of the chi-square distribution is as follows:

(D.15)

$$f(X) = \frac{X^{(k-2)/2}\exp(-X \big/ 2)}{2^{k/2}\Gamma(k \big/ 2)} \text{ for} X \text{ &gt; } 0$$
$$= 0 \text{ elsewhere}$$

where Γ(*k* / 2), is the **gamma function** with argument *k*/2, where *k* represents the *df* (see Figure D.3).

Figure D.3 The Chi-Square Distribution for Select Degrees of Freedom



D.3.1 Properties of the Chi-Square Distribution

1. The range of *X* is 0 ≤ *X* ≤ ∞.
2. It is a skewed distribution, the degree of the skewness depending on the *df*.
3. Its mean value is *k* and its variance is 2*k*, a unique property of this distribution.
4. Its mode is *k* − 2, *k* &gt; 2.
5. Its median is *k* − 2/3 (approximately for large *k*).
6. Its coefficient of skewness is $2^{-3/2}K^{-1/2}$.
7. Its coefficient of kurtosis is 3 + 12/*k*.
8. As the *df* increases, the chi-square distribution becomes increasingly symmetrical. As a matter of fact, for *df* in excess of 100, the variable
   (D.16)

$$\sqrt{2\chi^2} - \sqrt{(2k-1)} \sim N(0, 1)$$

can be treated as a standardized normal variable, where *k* is the *df*.

9. If Z_1 and Z_2 are two independent chi-square variables, with k_1 and k_2*df*, respectively, then the sum $Z_1 + Z_2$ is also a chi-square variable with $df = k_1 + k_2$.

D.4 Student's *t* Distribution

If Z_1 is a standardized normal variable [ie., $Z_1 \sim N(0,1)$] and another variable Z_2 follows the chi-square distribution with *k df* and is independent of Z_1, then

(D.17)

$$t = \frac{Z_1}{\sqrt{Z_2 / k}}$$
$$= \frac{Z_1\sqrt{k}}{\sqrt{Z_2}} \sim t_k$$

A test based on the *t* test is called (Student's) ***t* test**.

The properties of the *t* distribution are as follows:

1. The *t* distribution, like the normal distribution, is symmetrical, but it is flatter than the normal distribution. But as the *df* increases, the *t* distribution approximates the normal distribution. The approximation is reasonable for *k* &gt; 30.
2. The mean of the *t* distribution is zero, and its variance is *k*/(*k* − 2), which exists if *k* &gt; 2.
3. Coefficient of skewness = 0, *k* &gt; 3.
4. Coefficient of kurtosis = 3(*k* − 2)/(*k* − 4), *k* &gt; 4.
5. If $X_1, X_2, \ldots, X_n$ are iid $N(\mu,\sigma^2)$ random variables, then

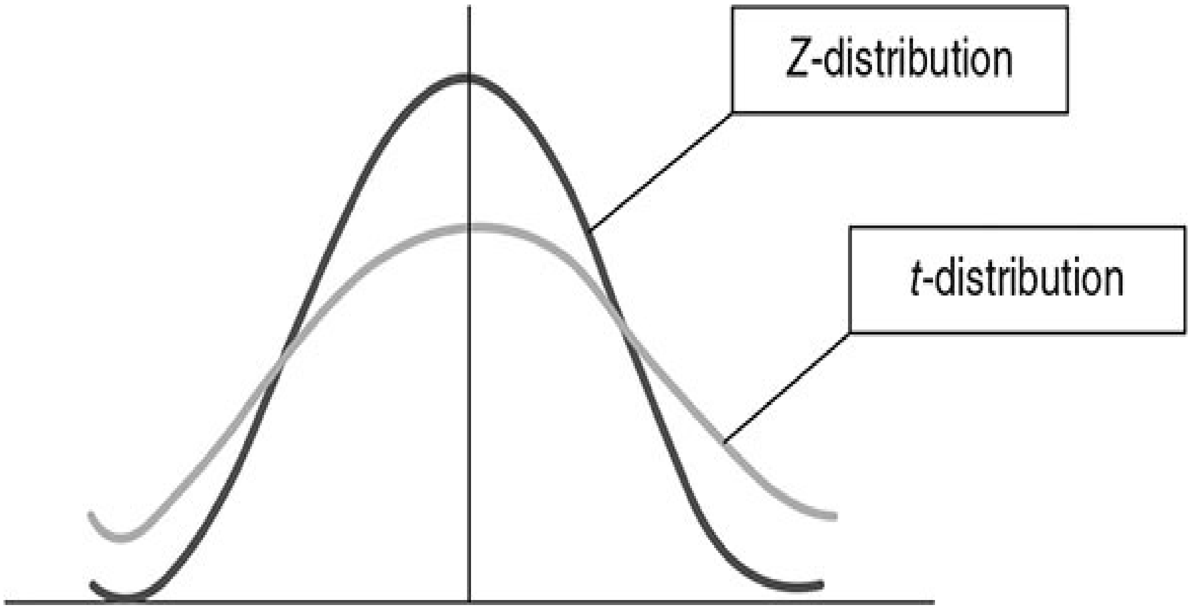
   (D.18)

   $$\frac{\overline{X} - \mu}{s}\sqrt{n} \sim t_{n-1}$$

where S^2 = sample variance.

© 2019 by DNG Econometrics, Inc

Figure D.4 The Normal and t Distributions



6. The *t* distribution with 1 *df* is called the Cauchy distribution (see Figure D.4).

D.5 Fisher's *F* Distribution

If Z_1 and Z_2 are independently distributed chi-square variables with k_1 and k_2*df*, respectively, the variable

(D.19)

$$F = \frac{Z_1 / k_1}{Z_2 / k_2} \sim F_{k_1 k_2}$$

follows the *F* distribution (see Figure D.5).

Note: k_1 is called the *numerator df* and k_2 is called the *denominator df.*

A test based on the *F* distribution is called an ***F* test**.

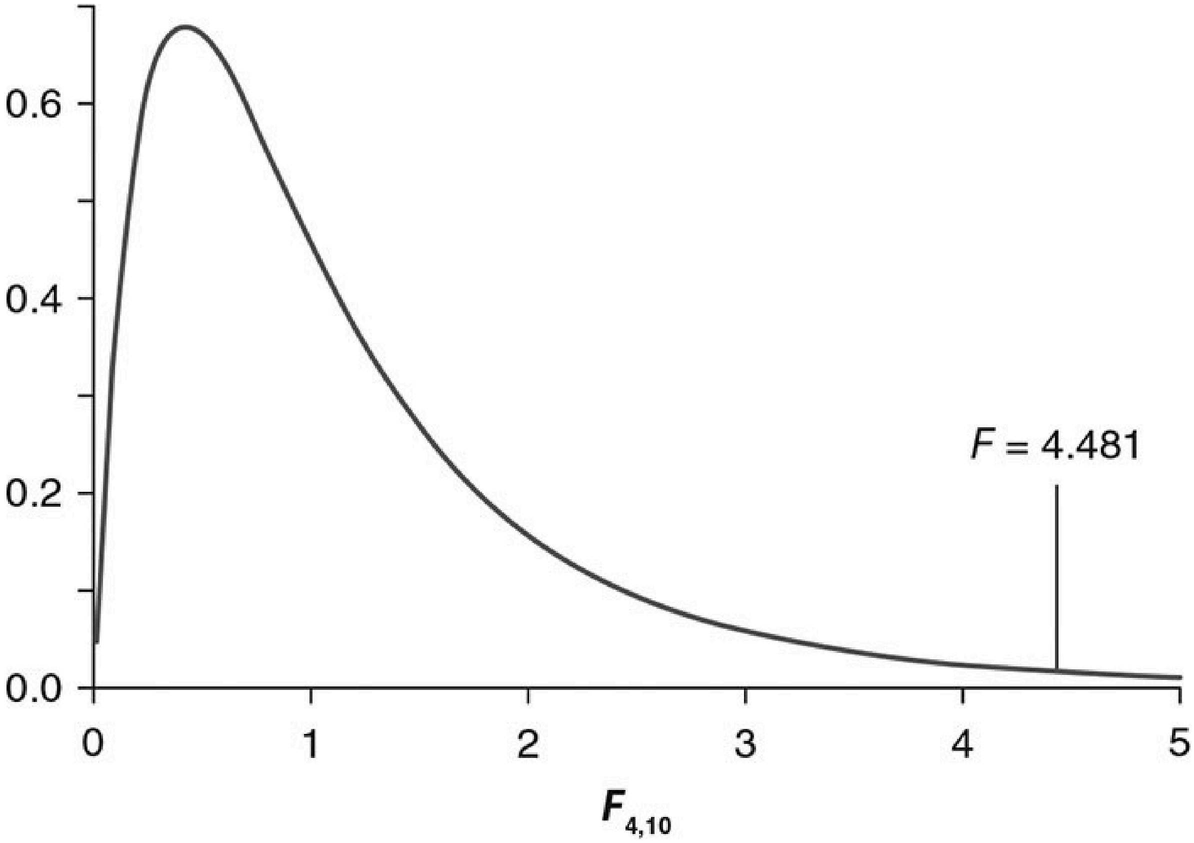
D.5.1 Properties of the F Distribution

1. Like the chi-square distribution, the *F* distribution is skewed to the right. But it can be shown that as k_1 and k_2 become increasingly larger, the *F* distribution approaches the normal

© 2019 by DNG Econometrics, Inc

distribution.

Figure D.5 The F Distribution for 4 and 10 Degrees of Freedom



2. The mean value of an *F*-distributed variable is $k2/(k2 - 2)$, which is defined for $k2$ &gt; 2, and its variance is

(D.20)

$$\frac{2k_2^2(k_1 + k_2 - 2)}{k_1(k_2 - 2)^2(k_2 - 4)}$$

which is defined for $k2$ &gt; 4.

3. Coefficient of skewness is

(D.21)

$$\frac{(2k_1 + k_2 - 2)[8(k_2 - 4)]^{1/2}}{k_1^{1/2}(k_2 - 6)(k_1 + k_2 - 2)^{1/2}} \quad k_2 \text{ &gt; } 6$$

4. Coefficient of kurtosis is

(D.22)

$$3 + \frac{12[(k_2 - 2)^2(k_2 - 4) + k_1(k_1 + k_2 - 2)(5k_2 - 22)]}{k_1(k_2 - 6)(k_2 - 8)(k_1 + k_2 - 2)} \quad \text{for} k_2 \text{ \&gt; } 8$$

5. The square of a *t*-distributed random variable with *k df* has an *F* distribution with 1 and *k df*. That is,

(D.23)

$$t_k^2 = F_{1,\, k}$$

6. If the denominator *df* $k2$ is fairly large, then we have the following relationship:

(D.24)

$$k_1 F \sim \chi_{k1}^2$$

In words, for a fairly large denominator *df*, the numerator *df* times the *F* value is approximately the same as a chi-square value with the numerator *df*.

7. If s_1^2 and s_2^2 are the variances of independent random samples of size $n1$ and $n2$ from normal populations with variances σ_1^2 and σ_2^2, then

(D.25)

$$F = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} \sim F_{n_1 - 1, n_2 - 1}$$

D.6 Relationships Among Probability Distributions

(D.26)

$$F_{1,\, k} = t_k^2$$

That is, the square of the *t* statistic with *k df* is equal to an *F* statistic with 1 *df* in the numerator and *k df* in the denominator.

(D.27)

$$mF_{m,\, n} = \chi_m^2 \quad n \to \infty$$

That is, for large denominator *df*, the numerator *df* times the *F* value is approximately equal to the chi-square value with the numerator *df*, where *m* is the numerator *df* and *n* is the denominator *df*.

© 2019 by DNG Econometrics, Inc

(D.28)

$$Z = \sqrt{2\chi^2} - \sqrt{2k - 1} \sim N(0,\ 1)$$

This states that for sufficiently large *df*, the chi-square distribution can be approximated by the standard normal distribution, where *k* is *df*.

D.7 Uniform Distributions

There are two types of uniform distributions: (1) discrete uniform and (2) continuous uniform.

D.7.1 Discrete Uniform Distribution

Let *n* &gt; 1 be an integer. Suppose the variable ***X*** has mass function given by

(D.29)

$$\begin{aligned} f(\boldsymbol{X}) &= 1 / n \text{ for } \boldsymbol{X} = 1,\ 2,\ \ldots,\ n \\ &= 0 \text{ otherwise} \end{aligned}$$

We say that ***X*** has a uniform distribution on (1, 2, …, *n*) (see Figure D.6).

(D.30)

$$\text{Mean of } X = (n + 1) / 2$$

(D.31)

$$\text{Variance of } X = (n^{2-1}) / 12$$

D.7.2 Continuous Uniform Distribution

The continuous random variable ***X*** has a uniform distribution over the interval (*a*, *b*) if its density function is given by

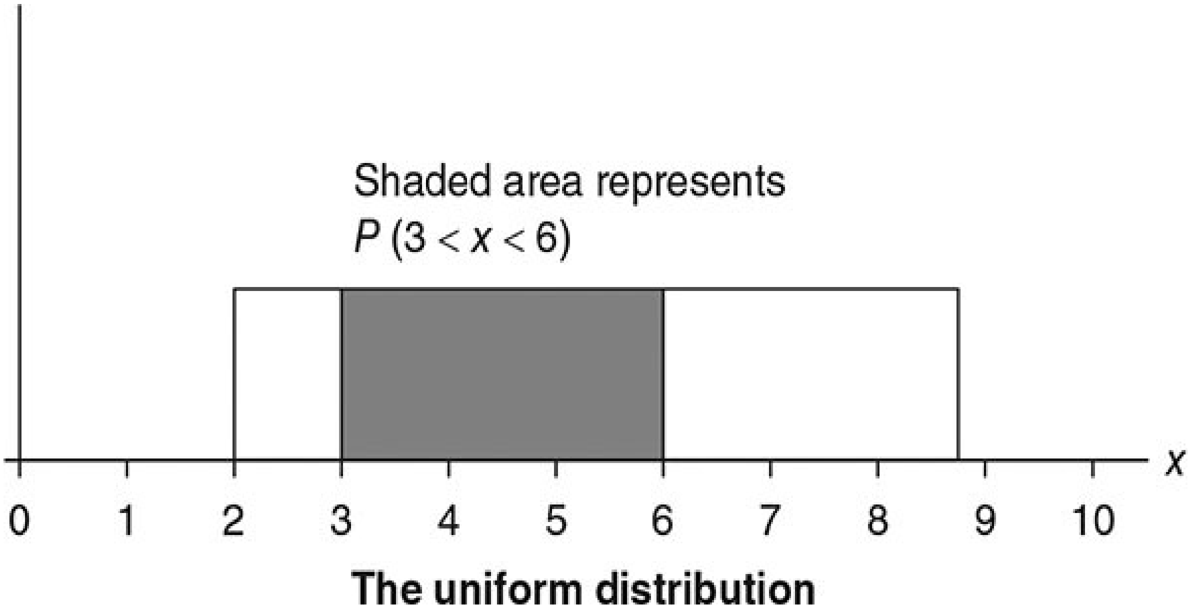
(D.32)

$$f(\boldsymbol{X}) = \begin{cases} \dfrac{1}{b - a} \ a \text{ \&lt; } x \text{ \&lt; } b \end{cases}$$

$$= 0 \text{ otherwise}$$

© 2019 by DNG Econometrics, Inc

Figure D.6 Discrete Uniform Distribution



$$F(\boldsymbol{X}) = \begin{cases} 0 & \text{for} x \text{ &lt; } a \\ \frac{x-a}{b-a} & \text{for} a \le x \text{ &lt; } b \\ 1 & \text{for} x \ge b \end{cases}$$

(D.33)
mean = $(a + b)/2$

(D.34)
variance = $(b - a)^2 / 12$

D.8 Some Special Features of the Normal Distribution[3]

If $X_1, X_2, \ldots, X_n$ is a sample from the normal population with mean = μ and variance = σ^2 that is, $X_i \sim N(\mu, \sigma^2)$, then the following features apply:

1. The sample mean $\overline{X}$ and the sample variance S^2 are independent random variables.
2. $\overline{X} \sim N(\mu, \sigma^2 / n)$.

© 2019 by DNG Econometrics, Inc

3. $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$.

The proof of Statement 1 is slightly complicated (see references in Footnote 1). So for now, we will assume that this is the case.

Statement 2 is easy to establish:

(D.35)

$$E(\overline{X}) = \frac{1}{n}E\left(\sum_{1}^{n} X_i\right) = \frac{1}{n}(n \cdot \mu) = \mu$$

since the expectation of each X_i is μ.

(D.36)

$$\begin{aligned}\operatorname{var}\overline{X} &= \frac{1}{n^2}\operatorname{var}(X_1, X_2, \ldots, X_n)\\ &= \frac{1}{n^2}(n \cdot \sigma^2) = \frac{\sigma^2}{n}\end{aligned}$$

since the Xs are independently distributed, each with variance of σ^2.

To establish Statement 3 above, we start with the following identity:

(D.37)

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2$$

which we can express as

(D.38)

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2}$$

which is equivalent to

(D.39)

$$\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left[\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}\right]^2$$

Noting that $\left(\frac{X_i - \mu}{\sigma}\right)$ is a unit normal variable so its square is a chi-square random variable with 1 *df* and since

© 2019 by DNG Econometrics, Inc

the left-hand side of Equation (D.39) is the sum of n independent chi-square random variables, it has n *df*. The last term in Equation (D.39) is also a chi-square variable with 1 *df*.

Now since $\bar{X}$ and S^2 are independent by Statement 1 above, it follows the two terms on the right-hand side of Equation (D.39) are independent, hence the conclusion that $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with (n – 1) *df*.[4] Recall that the sum of two or more independent chi-square variables is also the chi-square variable with degrees of freedom equal to the sum of the degrees of freedom of associated chi-square variables.

4.

(D.40)

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

5. If $X_i \sim N(\mu_x, \sigma_x^2)$ and $Y_i \sim N(\mu_y, \sigma_y^2)$ and Xi and Yi are independent of each other, then

(D.41)

$$\frac{S_x 2 \,/\, S_y^2}{\sigma_x^2 \,/\, \sigma_y^2} \sim F_{(n,\, m)}$$

where S_x^2 and S_y^2 are the sample variances of X and Y and n and m are their associated degrees of freedom.

The proofs of Equations (D.40) and (D.41) can be found in the references cited in Footnotes 1 and 4.

Notes

[1] For derivations of the distributions discussed in this appendix, see any textbook in mathematical statistics. For example, see Hogg, R. V., Mckean, J., &amp; Craig, A. T. (2012). *Introduction to mathematical statistics* (7th ed.). Harlow, England: Pearson Education.

[2] For an application of various statistical tests, see Kanji, G. K. (1999). *100 statistical tests* (New ed.). London, England: Sage.

[3] The proofs of some of the following statements are rather involved. See Casella, G., &amp; Berger, R. L. (2000). *Statistical inference* (2nd ed.) Belmont, CA: Wadsworth; Rice, J. A. (2007). *Mathematical statistics and data analysis* (3rd ed.). Pacific Grove, CA: Brooks/Cole; Wasserman, L. (2005). *All of statistics: A concise course in statistical inference*. New York, NY: Springer Science &amp; Business Media.

© 2019 by DNG Econometrics, Inc

[4] For a rigorous proof of this statement, see Casella, G., &amp; Berger, R. L. (2002). *Statistical inference* (2nd ed., pp. 218–219). Belmont, CA: Wadsworth.