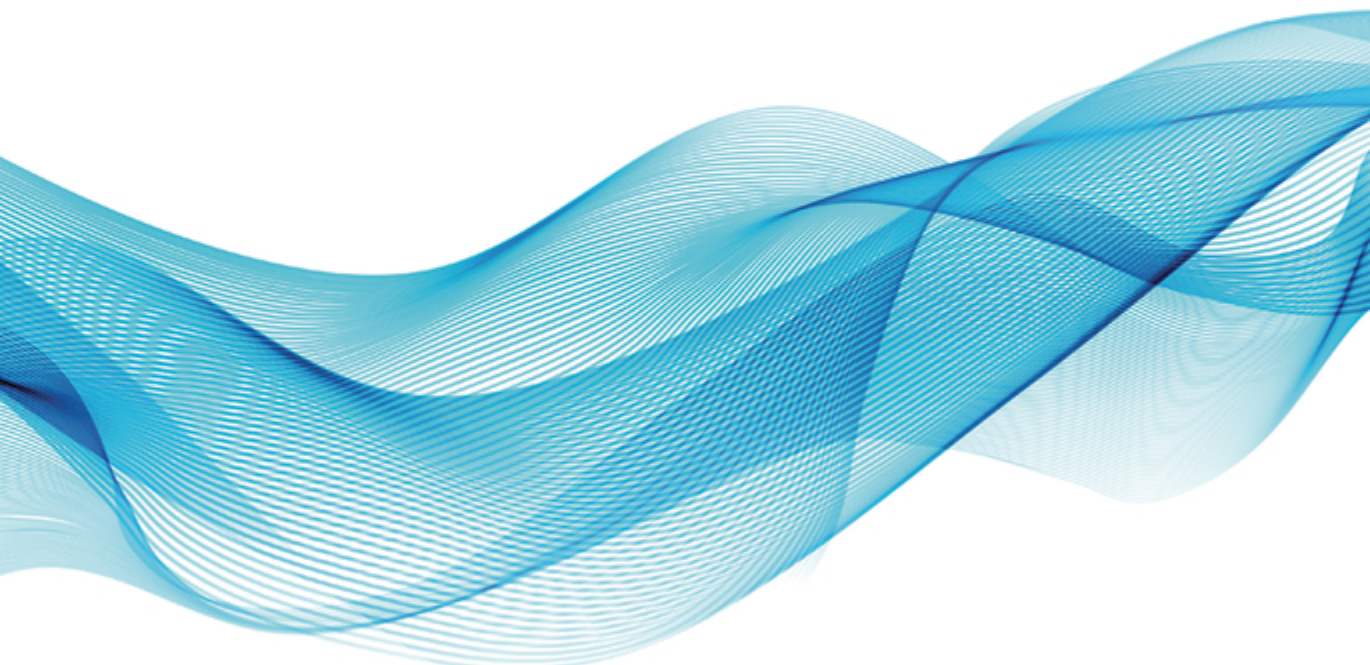
OXFORD

time series and panel data econometrics



M. HASHEM PESARAN

TIME SERIES AND PANEL DATA ECONOMETRICS

Time Series and Panel Data Econometrics

M. HASHEM PESARAN

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2015

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2015936093

ISBN 978–0–19–873691–2 (HB)
978–0–19–875998–0 (PB)

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

To my wife and in memory of my parents.

Preface

This book is concerned with recent developments in time series and panel data techniques for the analysis of macroeconomic and financial data. It provides a rigorous, nevertheless user-friendly, account of the time series techniques dealing with univariate and multivariate time series models, as well as panel data models. An overview of econometrics as a subject is provided in Pesaran (1987a) and updated in Geweke, Horowitz, and Pesaran (2008).

It is distinct from other time series texts in the sense that it also covers panel data models and attempts at a more coherent integration of time series, multivariate analysis, and panel data models. It builds on the author's extensive research in the areas of time series and panel data analysis and covers a wide variety of topics in one volume. Different parts of the book can be used as teaching material for a variety of courses in econometrics. It can also be used as a reference manual.

It begins with an overview of basic econometric and statistical techniques and provides an account of stochastic processes, univariate and multivariate time series, tests for unit roots, cointegration, impulse response analysis, autoregressive conditional heteroskedasticity models, simultaneous equation models, vector autoregressions, causality, forecasting, multivariate volatility models, panel data models, aggregation and global vector autoregressive models (*GVAR*). The techniques are illustrated using Microfit 5 (Pesaran and Pesaran (2009)) with applications to real output, inflation, interest rates, exchange rates, and stock prices.

The book assumes that the reader has done an introductory econometrics course. It begins with an overview of the basic regression model, which is intended to be accessible to advanced undergraduates, and then deals with more advanced topics which are more demanding and suited to graduate students and other interested scholars.

The book is organized into six parts:

Part I: Chapters 1 to 7 present the classical linear regression model, describe estimation and statistical inference, and discuss the violation of the assumptions underlying the classical linear regression model. This part also includes an introduction to dynamic economic modelling, and ends with a chapter on predictability of asset returns.

Part II: Chapters 8 to 11 deal with asymptotic theory and present the maximum likelihood and generalized method of moments estimation frameworks.

Part III: Chapters 12 and 13 provide an introduction to stochastic processes and spectral density analysis.

Part IV: Chapters 14 to 18 focus on univariate time series models and cover stationary *ARMA* models, unit root processes, trend and cycle decomposition, forecasting and univariate volatility models.

Part V: Chapters 19 to 25 consider a variety of reduced form and structural multivariate models, rational expectations models, as well as VARs, vector error corrections, cointegrating *VAR*s, *VARX* models, impulse response analysis, and multivariate volatility models.

Part VI: Chapters 26 to 33 considers panel data models both when the time dimension (T) of the panels is short, as well as when panels with N (the cross-section dimension) and T are large. These chapters cover a wide range of panel data models, starting with static panels with homogenous slopes and graduating to dynamic panels with slope heterogeneity, error cross-section dependence, unit roots, and cointegration.

There are also chapters dealing with the aggregation of large dynamic panels and the theory and practice of *GVAR* modelling. This part of the book focuses more on large N and T panels which are less covered in other texts, and draws heavily on my research in this area over the past 20 years starting with Pesaran and Smith (1995).

Appendices A and B present background material on matrix algebra, probability and distribution theory, and Appendix C provides an overview of Bayesian analysis.

This book has evolved over many years of teaching and research and brings together in one place a diverse set of research areas that have interested me. It is hoped that it will also be of interest to others. I have used some of the chapters in my teaching of postgraduate students at Cambridge University, University of Southern California, UCLA, and University of Pennsylvania. Undergraduate students at Cambridge University have also been exposed to some of the introductory material in Part I of the book. It is impossible to name all those who have helped me with the preparation of this volume. But I would like particularly to name two of my Cambridge Ph.D. students, Alexander Chudik and Elisa Tosetti, for their extensive help, particularly with the material in Part VI of the book.

The book draws heavily from my published and unpublished research. In particular:

Chapter 7 is based on Pesaran (2010).

Chapter 25 draws from Pesaran and Pesaran (2010).

Chapter 32 is based on Pesaran (2003) and Pesaran and Chudik (2014) where additional technical details and proofs are provided.

Chapter 31 is based on Breitung and Pesaran (2008) and provides some updates and extensions.

Chapter 33 is based on Chudik and Pesaran (2015b).

I would also like to acknowledge all my coauthors whose work has been reviewed in this volume. In particular, I would like to acknowledge Ron Smith, Bahram Pesaran, Allan Timmermann, Kevin Lee, Yongcheol Shin, Vanessa Smith, Cheng Hsiao, Michael Binder, Richard Smith, Alexander Chudik, Takashi Yamagata, Tony Garratt, Til Schermann, Filippo di Mauro, Stéphane Dées, Alessandro Rebucci, Adrian Pagan, Aman Ullah, and Martin Weale. It goes without saying that none of them is responsible for the material presented in this volume.

Finally, I would like to acknowledge the helpful and constructive comments and suggestions from two anonymous referees which provided me with further impetus to extend the coverage of the material included in the book and to improve its exposition over the past six months. Ron Smith has also provided me with detailed comments and suggestions over a number of successive drafts. I am indebted to him for helping me to see the wood from the trees over the many years that we have collaborated with each other.

Hashem Pesaran

Cambridge and Los Angeles
January 2015

Contents

List of Figures

List of Tables

Part I

Introduction to Econometrics

1 Relationship Between Two Variables

1.1 Introduction

There are a number of ways that a regression between two or more variables can be motivated. It can, for example, arise because we know *a priori* that there exists an exact linear relationship between Y and X, with Y being observed with measurement errors. Alternatively, it could arise if (Y, X) have a bivariate distribution and we are interested in the conditional expectations of Y given X, namely $E(Y \mid X)$, which will be a linear function of X either if the underlying relationship between Y and X is linear, or if Y and X have a bivariate normal distribution. A regression line can also be considered without any underlying statistical model, just as a method of fitting a line to a scatter of points in a two-dimensional space.

1.2 The curve fitting approach

We first consider the problem of regression purely as an act of fitting a line to a scatter diagram. Suppose that T pairs of observations on the variables Y and X, given by (y_1, x_1), (y_2, x_2), ..., (y_T, x_T), are available. We are interested in obtaining the equation of a straight line such that, for each observation x_t, the corresponding value of Y on a straight line in the (Y, X) plane is as 'close' as possible to the observed values y_t.

Immediately, different criteria of 'closeness' or 'fit' present themselves. Two basic issues are involved:

A: How to define and measure the *distance* of the points in the scatter diagram from the fitted line. There are three plausible ways to measure the distance of a point from the fitted line:

(i) perpendicular to x-axis
(ii) perpendicular to y-axis
(iii) perpendicular to the fitted line.

B: How to *add up* all such distances of the sampled observations. Possible weighting (adding-up) schemes are:

(i) simple average of the square of distances
(ii) simple average of the absolute value of distances
(iii) weighted averages either of squared distance measure or absolute distance measures.

The simplest is the combination **A**(i) and **B**(i), which gives the ordinary least squares (*OLS*) estimates of the regression of Y on X. The method of ordinary least squares will be extensively treated in the rest of this Chapter and in Chapter 2. The difference between **A**(i) and **A**(ii) can also be characterized as to which of the two variables, X or Y, is represented on the horizontal axis. The combination **A**(ii) and **B**(i) is also referred to as the 'reverse regression of Y on X'. Other combinations of distance/weighting schemes can also be considered. For example **A**(iii) and **B**(i) is called orthogonal regression, **A**(i) and **B**(ii) yields the absolute minimum distance regression. **A**(i) and **B**(iii) gives the weighted (or absolute distance) least squares (or absolute distance) regression.

1.3 The method of ordinary least squares

Treating X as the regressor and Y as the regressand, then choosing the distance measure, $d_t = \left|y_t - \alpha - \beta x_t\right|$, the *least squares criterion* function to be minimized is[1]

$$Q(\alpha, \beta) = \sum_{t=1}^{T} d_t^2 = \sum_{t=1}^{T} \left(y_t - \alpha - \beta x_t\right)^2.$$

The necessary conditions for this minimization problem are given by

$$\frac{\partial Q(\alpha, \beta)}{\partial \alpha} = \sum_{t=1}^{T} (-2)\left(y_t - \hat{\alpha} - \hat{\beta} x_t\right) = 0, \tag{1.1}$$

$$\frac{\partial Q(\alpha, \beta)}{\partial \beta} = \sum_{t=1}^{T} (-2x_t)\left(y_t - \hat{\alpha} - \hat{\beta} x_t\right) = 0. \tag{1.2}$$

Equations (1.1) and (1.1) are called *normal equations* for the *OLS* problem and can be written as

$$\sum_{t=1}^{T} \hat{u}_t = 0, \tag{1.3}$$

$$\sum_{t=1}^{T} \hat{u}_t x_t = 0, \tag{1.4}$$

[1] The notations $\sum_{t=1}^{T}$ and $\sum_t$ are used later to denote the sum of the terms after the summation sign over $t = 1, 2, \ldots, T$.

where

$$\hat{u}_t = y_t - \hat{\alpha} - \hat{\beta}x_t, \tag{1.5}$$

are the *OLS* residuals. The condition $\sum_{t=1}^{T} \hat{u}_t = 0$ also gives $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$, where $\bar{x} = \sum_{t=1}^{T} x_t/T$ and $\bar{y} = \sum_{t=1}^{T} y_t/T$, and demonstrates that the least squares regression line $\hat{y}_t = \hat{\alpha} + \hat{\beta}x_t$, goes through the sample means of Y and X. Solving (1.3) and (1.4) for $\hat{\beta}$, and hence for $\hat{\alpha}$, we have

$$\hat{\beta} = \frac{\sum_{t=1}^{T} x_t y_t - T\bar{x}\bar{y}}{\sum_{t=1}^{T} x_t^2 - T\bar{x}^2}, \tag{1.6}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{1.7}$$

or since

$$\sum_{t=1}^{T} (x_t - \bar{x})\left(y_t - \bar{y}\right) = \sum_{t=1}^{T} x_t y_t - T\bar{x}\bar{y},$$

$$\sum_{t=1}^{T} (x_t - \bar{x})^2 = \sum_{t=1}^{T} x_t^2 - T\bar{x}^2,$$

equivalently

$$\hat{\beta} = \frac{\sum_{t=1}^{T} (x_t - \bar{x})\left(y_t - \bar{y}\right)}{\sum_{t=1}^{T} (x_t - \bar{x})^2} = \frac{S_{XY}}{S_{XX}},$$

where

$$S_{XY} = \frac{\sum_{t=1}^{T} (x_t - \bar{x})\left(y_t - \bar{y}\right)}{T} = S_{YX},$$

$$S_{XX} = \frac{\sum_{t=1}^{T} (x_t - \bar{x})^2}{T}.$$

1.4 Correlation coefficients between *Y* and *X*

There are many measures of quantifying the strength of correlation between two variables. The most popular one is the product moment correlation coefficient which was developed by Karl Pearson and builds on an earlier contribution by Francis Galton. Other measures of correlations include the Spearman rank correlation and Kendall's τ correlation. We now consider each of these measures in turn and discuss their uses and relationships.

1.4.1 Pearson correlation coefficient

The Pearson correlation coefficient is a parametric measure of dependence between two variables, and assumes that the underlying bivariate distribution from which the observations are drawn have moments. For the variables Y and X, and the T pairs of observations $\{(y_1,x_1), (y_2,x_2),\ldots,(y_T,x_T)\}$ on these variables, Pearson or the simple correlation coefficient between Y and X is defined by

$$\hat{\rho}_{YX} = \frac{\sum_{t=1}^{T}(x_t-\bar{x})\left(y_t-\bar{y}\right)}{\left[\sum_{t=1}^{T}(x_t-\bar{x})^2\sum_{t=1}^{T}\left(y_t-\bar{y}\right)\right]^{1/2}} = \frac{S_{XY}}{(S_{YY}S_{XX})^{\frac{1}{2}}}, \tag{1.8}$$

It is easily seen that $\hat{\rho}_{YX}$ lies between -1 and $+1$. Notice also that the correlation coefficient between Y and X is the same as the correlation coefficient between X and Y, namely $\hat{\rho}_{XY} = \hat{\rho}_{YX}$. In this bivariate case we have the following interesting relationship between $\hat{\rho}_{XY}$ and the regression coefficients of the regression Y on X and the 'reverse' regression of X on Y. Denoting these two regression coefficients respectively by $\hat{\beta}_{Y\cdot X}$ and $\hat{\beta}_{X\cdot Y}$, we have

$$\hat{\beta}_{Y\cdot X}\hat{\beta}_{X\cdot Y} = \frac{S_{YX}S_{XY}}{(S_{XX}S_{YY})} = \hat{\rho}_{YX}^2. \tag{1.9}$$

Hence, if $\hat{\beta}_{Y\cdot X} > 0$ then $\hat{\beta}_{X\cdot Y} > 0$. Since $\hat{\rho}_{XY}^2 \leq 1$, if we assume that $\hat{\beta}_{Y\cdot X} > 0$ it follows that $\hat{\beta}_{X\cdot Y} \leq \frac{1}{\hat{\beta}_{Y\cdot X}}$. If we further assume that $0 < \hat{\beta}_{Y\cdot X} < 1$, then $\hat{\beta}_{X\cdot Y} = \frac{\hat{\rho}_{XY}^2}{\hat{\beta}_{Y\cdot X}} > \hat{\rho}_{XY}^2$.

1.4.2 Rank correlation coefficients

Rank correlation is often used in situations where the available observations are in the form of cardinal numbers, or if they are not sufficiently precise. Rank correlations are also used to avoid undue influences from outlier (extreme tail) observations on the correlation analysis. A number of different rank correlations have been proposed in the literature. In what follows we focus on the two most prominent of these, namely Spearman's rank correlation and Kendall's τ correlation coefficient. A classic treatment of the subject can be found in Kendall and Gibbons (1990).

Spearman rank correlation

Consider the T pairs of observations $\left\{(y_t,x_t), \text{for } t = 1,2,\ldots,T\right\}$ and rank the observations on each of the variables y and x, in an ascending (or descending) order. Denote the rank of these ordered series by $1,2,\ldots,T$, so that the first observation in the ordered set takes the value of 1, the second takes the value of 2, etc. The Spearman rank correlation, r_s, between y and x is defined by

$$r_s = 1 - \frac{6\sum_{t=1}^{T} d_t^2}{T(T^2-1)}, \tag{1.10}$$

where

$$d_t = Rank(y_t : \mathbf{y}) - Rank(x_t : \mathbf{x}),$$

and $Rank(y_t : \mathbf{y})$ is equal to a number in the range [1 *to* T] determined by the size of y_t relative to the other $T-1$ values of $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$. Note also that by construction $\sum_{t=1}^{T} d_t = 0$, and that $\sum_{t=1}^{T} d_t^2$ can only take even integer values and has a mean equal to $(T^3 - T)/6$. Hence $E(r_s) = 0$. The Spearman rank correlation can also be computed as a simple correlation between $ry_t = Rank(y_t : \mathbf{y})$ and $rx_t = Rank(x_t : \mathbf{x})$. It is easily seen that

$$r_s = \frac{\sum_{t=1}^{T}(ry_t - \overline{ry})(rx_t - \overline{rx})}{\left[\sum_{t=1}^{T}(ry_t - \overline{ry})^2\right]^{1/2}\left[\sum_{t=1}^{T}(rx_t - \overline{rx})^2\right]^{1/2}},$$

where

$$\overline{ry} = \overline{rx} = T^{-1}\sum_{t=1}^{T} ry_t = T^{-1}\sum_{t=1}^{T} rx_t = \frac{T+1}{2}.$$

Kendall's τ correlation

Another rank correlation coefficient was introduced by Kendall (1938). Consider the T pairs of ranked observations (ry_t, rx_t), associated with the quantitative measures (y_t, x_t), for $t = 1, 2, \ldots, T$ as discussed above. Then the two pairs of ranks (ry_t, rx_t) and (ry_s, rx_s) are said to be *concordant* if

$$(rx_t - rx_s)(ry_t - ry_s) > 0, \text{ concordant pairs for all } t \text{ and } s,$$

and *discordant* if

$$(rx_t - rx_s)(ry_t - ry_s) \leq 0, \text{ discordant pairs for all } t \text{ and } s.$$

Denoting the number of concordant pairs by P_T and the number of discordant pairs by Q_T, Kendall's τ correlation coefficient is defined by

$$\tau_T = \frac{2}{T(T-1)}(P_T - Q_T). \tag{1.11}$$

More formally

$$P_T = \sum_{t,s=1}^{T} I\left[(rx_t - rx_s)(ry_t - ry_s)\right],$$
$$Q_T = \sum_{t,s=1}^{T} I\left[-(rx_t - rx_s)(ry_t - ry_s)\right],$$

where $I(A) = 1$ if $A > 0$, and zero otherwise.

1.4.3 Relationships between Pearson, Spearman, and Kendall correlation coefficients

In the case where (y_t, x_t) are draws from a normal distribution we have

$$E(\tau_T) = \frac{2}{\pi}\sin^{-1}(\rho),$$

where ρ is the simple (Pearson) correlation coefficient between y_t and x_t. Furthermore,

$$E(r_s) = \rho_s + \frac{3(\tau - \rho_s)}{T+1},$$

where ρ_s is the population value of Spearman rank correlation. Finally, in the bivariate normal case we have

$$\rho = 2\sin\left(\frac{\pi\rho_s}{6}\right).$$

These relationships suggest the following indirect possibilities for estimation of the simple correlation coefficient, namely

$$\hat{\rho}_1 = \sin\left(\frac{\pi}{2}\tau_T\right),$$
$$\hat{\rho}_2 = 2\sin\left[\frac{\pi}{6}\left(r_s - \frac{3(\tau_T - r_s)}{T+1}\right)\right],$$

as possible alternatives to $\hat{\rho}$, the simple correlation coefficient. See Kendall and Gibbons (1990, p. 169). The alternative estimators, $\hat{\rho}_1$ and $\hat{\rho}_2$, are likely to have some merit over $\hat{\rho}$ in small samples in cases where the population distribution of (y_t, x_t) differs from bivariate normal and/or when the observations are subject to measurement errors.

Tests based on the different correlation measures are discussed in Section 3.4.

1.5 Decomposition of the variance of *Y*

It is possible to divide the total variation of Y into two parts, the variation of the estimated Y and a residual variation. In particular

$$\begin{aligned}\sum_{t=1}^{T}(y_t - \bar{y})^2 &= \sum_{t=1}^{T}\left[(\hat{y}_t - \bar{y}) - (\hat{y}_t - y_t)\right]^2 \\ &= \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{T}(\hat{y}_t - y_t)^2 - 2\sum_{t=1}^{T}(\hat{y}_t - y_t)(\hat{y}_t - \bar{y}) \\ &= \sum_{t=1}^{T}(y_t - \hat{y}_t)^2 + \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 + 2\sum_{t=1}^{T}\hat{u}_t(\hat{y}_t - \bar{y}).\end{aligned}$$

But, notice that

$$\sum_{t=1}^{T} \hat{u}_t \left(\hat{y}_t - \bar{y}\right) = \sum_{t=1}^{T} \hat{u}_t \left(\hat{\alpha} + \hat{\beta} x_t\right) - \sum_{t=1}^{T} \hat{u}_t \bar{y}$$
$$= \hat{\alpha} \sum_{t=1}^{T} \hat{u}_t + \hat{\beta} \sum_{t=1}^{T} \hat{u}_t x_t - \bar{y} \sum_{t=1}^{T} \hat{u}_t = 0,$$

since from the normal equations (1.3) and (1.4), $\sum_{t=1}^{T} \hat{u}_t = 0$ and $\sum_{t=1}^{T} \hat{u}_t x_t = 0$, then

$$\sum_{t=1}^{T} \left(y_t - y_t\right)^2 = \sum_{t=1}^{T} \left(\hat{y}_t - \bar{y}\right)^2 + \sum_{t=1}^{T} \left(y_t - \hat{y}_t\right)^2. \tag{1.12}$$

This decomposition of the total variations in Y forms the basis of the *analysis of variance*, which is described in the following table.

Source of variation	Sums of squares	Degrees of freedom	Mean square
Explained by the regression line	$\sum_{t=1}^{T} \left(\hat{y}_t - \bar{y}\right)^2$	2	$\frac{\sum_{t=1}^{T} (\hat{y}_t - \bar{y})^2}{2}$
Residual	$\sum_{t=1}^{T} \left(y_t - \hat{y}_t\right)^2$	$T-2$	$\frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{T-2}$
Total variation	$\sum_{t=1}^{T} \left(y_t - \bar{y}\right)^2$	T	$\frac{\sum_{t=1}^{T} (y_t - \bar{y})^2}{T}$

Proposition 1 highlights the relation between $\hat{\rho}_{XY}^2$ and the variance decomposition.

Proposition 1

$$\hat{\rho}_{XY}^2 = \frac{S_{XY}^2}{S_{XX} S_{YY}} = 1 - \frac{\sum_{t=1}^{T} \left(y_t - \hat{y}_t\right)^2}{\sum_{t=1}^{T} \left(y_t - \bar{y}\right)^2}. \tag{1.13}$$

Proof Notice that

$$1 - \frac{\sum_t \left(y_t - \hat{y}_t\right)^2}{\sum_t \left(y_t - \bar{y}\right)^2} = \frac{\sum_t \left(y_t - \bar{y}\right)^2 - \sum_t \left(y_t - \hat{y}_t\right)^2}{\sum_t \left(y_t - \bar{y}\right)^2},$$

and using the result in (1.12), we have

$$1 - \frac{\sum_t \left(y_t - \hat{y}_t\right)^2}{\sum_t \left(y_t - \bar{y}\right)^2} = \frac{\sum_t \left(\hat{y}_t - \bar{y}\right)^2}{\sum_t \left(y_t - \bar{y}\right)^2}.$$

Further, since $\hat{y}_t = \hat{\alpha} + \hat{\beta} x_t$, we have

$$\sum_t \left(\hat{y}_t - \bar{y}\right)^2 = \sum_t \left(\hat{\alpha} + \hat{\beta} x_t - \bar{y}\right)^2$$
$$= \sum_t \left[\hat{\beta}\left(x_t - \bar{x}\right) + \hat{\beta}\bar{x} + \hat{\alpha} - \bar{y}\right]^2,$$

By (1.1), $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x}$. Hence, it follows that

$$\sum_t \left(\hat{y}_t - \bar{y}\right)^2 = \hat{\beta}^2 \sum_t \left(x_t - \bar{x}\right)^2 = \frac{S_{XY}^2}{S_{XX}^2} \cdot S_{XX} = \frac{S_{XY}^2}{S_{XX}},$$
$$1 - \frac{\sum_t \left(y_t - \hat{y}_t\right)^2}{\sum_t \left(y_t - \bar{y}\right)^2} = \frac{S_{YY}^2}{S_{YY} S_{XX}} = \hat{\rho}_{XY}^2.$$

The above result is important since it also provides a natural generalization of the concept of the simple correlation coefficient, $\hat{\rho}_{XY}$, to the multivariate regression case, where it is referred to as the multiple correlation coefficient (see Section 2.10).

1.6 Linear statistical models

So far we have viewed the regression equation as a line fitted to a scatter of points in a two-dimensional space. As such it is purely a descriptive scheme that attempts to summarize the scatter of points by a single regression line. An alternative procedure would be to adopt a statistical model where the regression disturbances, u_t's, are characterized by a probability distribution. Under this framework there are two important statistical models that are used in the literature:

A: Classical linear regression model. This model assumes that the relationship between Y and X is a linear one:

$$y_t = \alpha + \beta x_t + u_t, \tag{1.14}$$

and that the disturbances u_ts satisfy the following assumptions:

(i) *Zero mean*: the disturbances u_t have zero means, i.e., $E(u_t) = 0$.

(ii) *Homoskedasticity:* conditional on x_t the disturbances u_t have constant conditional variance. $Var\left(u_t \mid x_s\right) = \sigma^2$, for all t and s.

(iii) *Non-autocorrelated error*: the disturbances u_t are serially uncorrelated. $\text{Cov}(u_t, u_s) = 0$ for all $t \neq s$.

(iv) *Orthogonality:* the disturbances u_t and the regressor x_t are uncorrelated, or conditional on x_s, u_t has a zero mean (namely $E\left(u_t \mid x_s\right) = 0$, for all t and s).

Assumption (i) ensures that the unconditional mean of y_t is correctly specified by the regression equation. The other assumptions can be relaxed and are introduced to provide a simple model that can be used as a benchmark in econometric analysis.

B: Another way of motivating the linear regression model is to focus on the *joint* distribution of Y and X, and assume that this distribution is normal with constant means, variances and covariances. In this case the regression of Y on X defined as the conditional mean of Y given a particular value of X, say $X = x$ will be a linear function of x. In particular we have:

$$E(Y|X = x_t) = \alpha + \beta x_t, \tag{1.15}$$

and

$$Var(Y|X = x_t) = Var(Y)\left(1 - \rho_{XY}^2\right), \tag{1.16}$$

and where $Var(Y)$ is the unconditional variance of Y and

$$\rho_{XY} = \text{Cov}(Y, X)/\sqrt{Var(X)\,Var(Y)}$$

is the population correlation coefficient between Y and X.

The parameters α and β are related to the moments of the joint distribution of Y and X in the following manner:

$$\alpha = E(Y) - \frac{\text{Cov}(X, Y)}{Var(X)}E(X), \tag{1.17}$$

and

$$\beta = \frac{\text{Cov}(X, Y)}{Var(X)} = \rho_{XY}\sqrt{\frac{Var(Y)}{Var(X)}}. \tag{1.18}$$

Using (1.17) and (1.18), relation (1.15) can also be written as:

$$E(Y|X = x_t) = E(Y) + \frac{\text{Cov}(X, Y)}{Var(X)}\left[x_t - E(X)\right]. \tag{1.19}$$

Model B does not postulate a linear relationship between Y and X, but assumes that (Y, X) have a bivariate normal distribution. In contrast, model A assumes linearity of the relationship between Y and X, but does not necessarily require that the joint probability distribution of (Y, X) be normal. It is clear that under assumption (iv), (1.14) implies (1.15). Also (1.15) can be used to obtain (1.14) by defining u_t to be

$$u_t = y_t - E(Y|X = x_t),$$

or more simply

$$u_t = y_t - E\left(y_t|x_t\right). \tag{1.20}$$

It is in the light of this expression that u_t's are also often referred to as 'innovations' or 'unexpected components' of y_t.

Both the above statistical models are used in the econometric literature. The two models can also be combined to yield the 'classical normal linear regression model' which adds the extra assumption that u_t are normally distributed to the list of the *four* basic assumptions of the classical linear regression model set out above.

Finally, it is worth noting that under the normality assumption using (1.16) we also have

$$Var\,(u_t\,|x_t\,) = \sigma^2 = Var\,(Y)\left(1 - \rho_{YX}^2\right). \tag{1.21}$$

Hence,

$$\rho_{YX}^2 = 1 - \frac{\sigma^2}{Var\,(Y)},$$

which is the population value of the sample correlation coefficient defined by (1.8) and (1.13).

1.7 Method of moments applied to bivariate regressions

The *OLS* estimators can also be motivated by the method of moments originally introduced by Karl Pearson in 1894. Under the method of moments the parameters α and β are estimated by replacing population moments by their sample counterparts. Under Assumptions (i) and (iv) above that the errors, u_t, have zero means and are orthogonal to the regressors, we have the following two moment conditions

$$\begin{aligned} E(u_t) &= E(y_t - \alpha - \beta x_t) = 0, \\ E(x_t u_t) &= E\left[x_t(y_t - \alpha - \beta x_t)\right] = 0, \end{aligned}$$

which can also be written equivalently as

$$\begin{aligned} E(y_t) &= \alpha + \beta E\,(x_t)\,, \\ E(y_t x_t) &= \alpha E(x_t) + \beta E(x_t^2). \end{aligned}$$

It is clear that α and β can now be derived in terms of the population moments, $E(y_t)$, $E(x_t)$, $E(x_t^2)$, and $E(y_t x_t)$, namely

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1 & E\,(x_t) \\ E\,(x_t) & E(x_t^2) \end{pmatrix}^{-1} \begin{pmatrix} E(y_t) \\ E(y_t x_t) \end{pmatrix}. \tag{1.22}$$

The inverse exists if $Var(x_t) = E(x_t^2) - \left[E\,(x_t)\right]^2 > 0$. *The method of moment estimators of* α and β are obtained when the population moments in the above expression are replaced by the sample moments which are given by

$$\begin{aligned} &\hat{E}(y_t) = \bar{y}, \hat{E}\,(x_t) = \bar{x}, \\ &\hat{E}(x_t^2) = T^{-1}\sum_{t=1}^{T} x_t^2, \hat{E}(y_t x_t) = T^{-1}\sum_{t=1}^{T} y_t x_t. \end{aligned}$$

Using these sample moments in (1.22) gives $\hat{\alpha}_{MM}$ and $\hat{\beta}_{MM}$, that are easily verified to be the same as the *OLS* estimators given by (1.7) and (1.6).

In cases where the number of moment conditions exceed the number of unknown parameters, the method of moments is generalized to take account of the additional moment conditions in an efficient manner. The resultant estimator is then referred to as the generalized method of moments (*GMM*), which will be discussed in some detail in Chapter 10.

1.8 The likelihood approach for the bivariate regression model

An alternative estimation approach developed by R. A. Fisher over the period 1912–22 (building on the early contributions of Gauss, Laplace, and Edgeworth) is to estimate the unknown parameters by maximizing their likelihood. The likelihood function is then given by the joint probability distribution of the observations. In the case of the bivariate classical regression model the likelihood is obtained from the joint distribution of $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$, conditional on $\mathbf{x} = (x_1, x_2, \ldots, x_T)'$. To obtain this joint probability distribution, in addition to the assumptions of the classical linear regression, (i)-(iv) given in Section 1.6, we also need to specify the probability distribution of the errors, u_t. Typically, it is assumed that $u_t's$ are normally distributed, and the joint probability distribution of $\mathbf{y}$ conditional on $\mathbf{x}$, is then obtained as (since the Jacobian of the transformation between y_t and u_t is unity)

$$\Pr\left(\mathbf{y}\,|\mathbf{x},\alpha,\beta,\sigma^2\right) = \Pr(u_1, u_2, \ldots, u_T\,|\mathbf{x}).$$

But under the assumption that the errors are normally distributed, the non-autocorrelated error assumption, (iii), implies that the errors are independently distributed and hence we have

$$\Pr\left(\mathbf{y}\,|\mathbf{x},\alpha,\beta,\sigma^2\right) = \Pr(u_1)\Pr(u_2)\ldots.\Pr(u_T).$$

But the probability density function of a $N(0,\sigma^2)$ random variable is given by

$$\Pr(u_t) = (2\pi\sigma^2)^{-1/2}\exp\left(\frac{-1}{2\sigma^2}u_t^2\right).$$

Using this result and noting that $u_t = y_t - \alpha - \beta x_t$, we have

$$\Pr\left(\mathbf{y}\,|\mathbf{x},\alpha,\beta,\sigma^2\right) = (2\pi\sigma^2)^{-T/2}\exp\left[\frac{-\sum_{t=1}^{T}\left(y_t-\alpha-\beta x_t\right)^2}{2\sigma^2}\right].$$

The likelihood of the unknown parameters, which we collect in the 3×1 vector $\boldsymbol{\theta} = (\alpha,\beta,\sigma^2)'$, is the same as the above joint density function, but is viewed as a function of $\boldsymbol{\theta}$ rather than $\mathbf{y}$. Denoting the likelihood function of $\boldsymbol{\theta}$ by $L_T(\boldsymbol{\theta})$ we have

$$L_T(\boldsymbol{\theta}) = (2\pi\sigma^2)^{-T/2}\exp\left[\frac{-\sum_{t=1}^{T}\left(y_t-\alpha-\beta x_t\right)^2}{2\sigma^2}\right]. \tag{1.23}$$

To obtain the maximum likelihood estimator (*MLE*) of $\boldsymbol{\theta}$ it is often more convenient to work with the logarithm of the likelihood function, referred to as the log-likelihood function, which we denote by $\ell_T(\boldsymbol{\theta})$. Using (1.23) we have

$$\ell_T(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi\sigma^2) - \frac{\sum_{t=1}^{T}\left(y_t - \alpha - \beta x_t\right)^2}{2\sigma^2}.$$

It is now clear that maximization of $\ell_T(\boldsymbol{\theta})$ with respect to α and β will be the same as minimizing $\sum_{t=1}^{T}\left(y_t - \alpha - \beta x_t\right)^2$ with respect to these parameters, which establish that the *MLE* of α and β is the same as their *OLS* estimators, namely $\hat{\alpha}_{ML} = \hat{\alpha}$, and $\hat{\beta}_{ML} = \hat{\beta}$, where $\hat{\alpha}$, and $\hat{\beta}$ are given by (1.7) and (1.6), respectively. The *MLE* of σ^2 can be obtained by taking the first derivative of $\ell_T(\boldsymbol{\theta})$ with respect to σ^2. We have

$$\frac{\partial \ell_T(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{\sum_{t=1}^{T}\left(y_t - \alpha - \beta x_t\right)^2}{2\sigma^4}.$$

Setting $\partial \ell_T(\boldsymbol{\theta})/\partial\sigma^2 = 0$ and solving for $\hat{\sigma}^2_{ML}$ in terms of the *MLE* of α and β now yields

$$\hat{\sigma}^2_{ML} = \frac{\sum_{t=1}^{T}\left(y_t - \hat{\alpha}_{ML} - \hat{\beta}_{ML}x_t\right)^2}{T} = \frac{\sum_{t=1}^{T}\left(y_t - \hat{\alpha} - \hat{\beta}x_t\right)^2}{T} = \frac{\sum_{t=1}^{T}\hat{u}_t^2}{T}, \qquad (1.24)$$

where $\hat{u}_t$ is the *OLS* residual, given be (1.5).

The likelihood approach is used extensively in subsequent chapters. For an analysis of the *MLE* for multiple regression models see (2.4). The general theory of maximum likelihood estimation is provided in Chapter 9.

1.9 Properties of the *OLS* estimators

Under the classical assumptions (i)–(iv) in Section 1.6 above, the *OLS* estimators of α and β possess the following properties:

(i) $\hat{\alpha}$ and $\hat{\beta}$ are *unbiased* estimators. Namely, that $E\left(\hat{\alpha}\right) = a$ and $E\left(\hat{\beta}\right) = \beta$, where α and β are the 'true' values of the regression coefficients.

(ii) Both estimators are *linear functions* of the values of y_t.

(iii) Among the class of linear unbiased estimators, $\hat{\alpha}$ and $\hat{\beta}$ have the *least variances*. This result is known as the *Gauss–Markov theorem*.

In what follows we present a proof of properties (i) to (iii) for $\hat{\beta}$. A similar proof can also be established for $\hat{\alpha}$. Recall that

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{t=1}^{T}\left(y_t - \bar{y}\right)\left(x_t - \bar{x}\right)}{\sum_{t=1}^{T}\left(x_t - \bar{x}\right)^2}.$$

But the numerator of this ratio can be written as

$$\sum_{t=1}^{T}\left(y_t-\bar{y}\right)\left(x_t-\bar{x}\right)=\sum_{t=1}^{T}y_t\left(x_t-\bar{x}\right)-\sum_{t=1}^{T}\bar{y}\left(x_t-\bar{x}\right),$$

and since $\sum_{t=1}^{T}\bar{y}\left(x_t-\bar{x}\right)=\bar{y}\sum_{t=1}^{T}\left(x_t-\bar{x}\right)=0$, then

$$\sum_{t=1}^{T}\left(y_t-\bar{y}\right)\left(x_t-\bar{x}\right)=\sum_{t=1}^{T}y_t\left(x_t-\bar{x}\right).$$

Hence $\hat{\beta}$ can be written as a weighted linear function of y_t's

$$\hat{\beta}=\sum_{t=1}^{T}w_t y_t, \tag{1.25}$$

where the weights

$$w_t=\frac{x_t-\bar{x}}{\sum_{t=1}^{T}\left(x_t-\bar{x}\right)^2} \tag{1.26}$$

are fixed and add up to zero, namely $\sum_{t=1}^{T}w_t=0$. This establishes property (ii).

Notice that x_t's are taken as given, which is justified if they are strictly exogenous. Further discussion of the concept of strict exogeneity is given in Section 2.2, but in the present context x_t will be strictly exogenous if it is uncorrelated with current, past, as well as future values of the error terms, u_s; more specifically if $Cov(x_t, u_s)=0$, for all values of t and s. Under this assumption, taking conditional expectations of both sides of (1.25), we have:

$$\begin{aligned}E\left(\hat{\beta}\right)&=E\left(\sum_{t=1}^{T}w_t y_t\,|x_1,x_2,\ldots,x_T\right)\\&=\sum_{t=1}^{T}w_t E\left(y_t\,|x_t\right),\end{aligned}$$

But using (1.14) or (1.15), conditional on x_t, we have $E\left(y_t\,|x_t\right)=\alpha+\beta x_t$. Consequently,

$$\begin{aligned}E\left(\hat{\beta}\right)&=\sum_{t=1}^{T}w_t\left(\alpha+\beta x_t\right)\\&=\alpha\sum_{t=1}^{T}w_t+\beta\sum_{t=1}^{T}w_t x_t.\end{aligned} \tag{1.27}$$

However, using (1.26) we have

$$\sum_{t=1}^{T} w_t x_t = \frac{\sum_{t=1}^{T} x_t \left(x_t - \bar{x}\right)}{\sum_{t=1}^{T} \left(x_t - \bar{x}\right)^2},$$

and since

$$\begin{aligned}
\sum_{t=1}^{T} \left(x_t - \bar{x}\right)^2 &= \sum_{t=1}^{T} \left(x_t - \bar{x}\right) \left(x_t - \bar{x}\right) \\
&= \sum_{t=1}^{T} x_t \left(x_t - \bar{x}\right) - \sum_{t=1}^{T} \bar{x} \left(x_t - \bar{x}\right) \\
&= \sum_{t=1}^{T} x_t \left(x_t - \bar{x}\right) - \bar{x} \sum_{t=1}^{T} \left(x_t - \bar{x}\right) \\
&= \sum_{t=1}^{T} x_t \left(x_t - \bar{x}\right),
\end{aligned}$$

it then follows that $\sum_{t=1}^{T} w_t x_t = 1$. We have also seen that $\sum_{t=1}^{T} w_t = 0$, hence it follows from (1.27) that $E\left(\hat{\beta}\right) = \beta$, which establishes that $\hat{\beta}$ is an unbiased estimator, that is, point (i).

The variance of $\hat{\beta}$ can also be computed easily using (1.25). We have

$$\begin{aligned}
Var\left(\hat{\beta}\right) &= \sum_{t=1}^{T} w_i^2 Var\left(y_t \,|x_t\right) \\
&= \sum_{t=1}^{T} w_i^2 Var\left(u_t \,|x_t\right) \\
&= \sigma^2 \sum_{t=1}^{T} w_i^2,
\end{aligned}$$

and using (1.26) yields

$$Var\left(\hat{\beta}\right) = \frac{\sigma^2}{\sum_{t=1}^{T} \left(x_t - \bar{x}\right)^2} = \frac{\sigma^2}{S_{XX}}. \tag{1.28}$$

Similarly, we have

$$Var\left(\hat{\alpha}\right) = \frac{\sigma^2 \sum_{t=1}^{T} x_t^2}{T \sum_{t=1}^{T} \left(x_t - \bar{x}\right)^2}, \tag{1.29}$$

and

$$\text{Cov}\left(\hat{\alpha},\hat{\beta}\right)=\frac{-\sigma^2\bar{x}}{\sum_{t=1}^{T}(x_t-\bar{x})^2}. \tag{1.30}$$

The Gauss–Markov theorem (i.e., property (iii) above) states that among all linear, unbiased estimators of β (or α) the *OLS* estimator, $\hat{\beta}$, has the smallest variance. To prove this result consider *another* linear unbiased estimator of β and denote it by $\tilde{\beta}$. Then by assumption

$$\tilde{\beta}=\sum_{t=1}^{T}\tilde{w}_t y_t,$$

where $\tilde{w}_t$ are fixed weights (which do not depend on y_t) and satisfy the conditions

$$\sum_{t=1}^{T}\tilde{w}_t=0, \tag{1.31}$$

and

$$\sum_{t=1}^{T}\tilde{w}_t x_t=1. \tag{1.32}$$

These two conditions ensure that $\tilde{\beta}$ is an unbiased estimator of β, that is, that $E\left(\tilde{\beta}\right)=\beta$. Suppose now $\tilde{w}_t$ differ from w_t, the *OLS* weights given in (1.26), by the amount δ_t and let

$$\tilde{w}_t=w_t+\delta_t,\qquad t=1,2,\ldots,T, \tag{1.33}$$

where δ_t is the amount of discrepancy between the two weighting schemes. Since $\sum_{t=1}^{T}w_t=\sum_{t=1}^{T}\tilde{w}_t=0$. It follows also that $\sum_{t=1}^{T}\delta_t=0$, and since $\sum_{t=1}^{T}w_t x_t=\sum_{t=1}^{T}\tilde{w}_t x_t=1$, then we should also have $\sum_{t=1}^{T}\delta_t x_t=0$.

The variance of $\tilde{\beta}$ is now given by

$$\begin{aligned}Var\left(\tilde{\beta}\right)&=\sum_{t=1}^{T}\tilde{w}_t^2 Var\left(y_t\,|x_t\right)\\&=\sigma^2\sum_{t=1}^{T}\tilde{w}_t^2,\end{aligned}$$

and using (1.33)

$$Var\left(\tilde{\beta}\right)=\sigma^2\left(\sum_{t=1}^{T}w_i^2+\sum_{t=1}^{T}\delta_t^2+2\sum_{t=1}^{T}w_t\delta_t\right).$$

But, using (1.26),

$$\sum_{t=1}^{T} w_t \delta_t = \frac{\sum_{t=1}^{T} \delta_t (x_t - \bar{x})}{\sum_{t=1}^{T} (x_t - \bar{x})^2}.$$

The numerator of this ratio can be written more fully as

$$\sum_{t=1}^{T} \delta_t (x_t - \bar{x}) = \sum_{t=1}^{T} \delta_t x_t - \bar{x} \sum_{t=1}^{T} \delta_t,$$

which is equal to zero. Recall that $\sum_{t=1}^{T} \delta_t = 0$, and $\sum_{t=1}^{T} \delta_t x_t = 0$. Hence $\sum_{t=1}^{T} w_t \delta_t = 0$, and

$$Var\left(\tilde{\beta}\right) = \sigma^2 \left\{ \sum_{t=1}^{T} w_i^2 + \sum_{t=1}^{T} \delta_t^2 \right\} \geq Var\left(\tilde{\beta}\right),$$

which establishes the Gauss–Markov theorem for $\hat{\beta}$. The equality sign holds if and only if $\delta_t = 0$ for all i. The proof of the Gauss–Markov theorem for the multivariate case is presented in Section 2.7.

1.9.1 Estimation of σ^2

Since $Var\left(\hat{\alpha}\right)$ and $Var\left(\tilde{\beta}\right)$ depend on the unknown parameter, σ^2 (the variance of the disturbance term), in order to obtain estimates of the variances of the *OLS* estimators, it is also necessary to obtain an estimate of σ^2. For this purpose we first note that

$$\sigma^2 = Var\left(u_t \,|x_t\right) = E\left(u_t^2\right).$$

It is, therefore, reasonable to interpret σ^2 as the mean value of the squared disturbances. A moment estimator of σ^2 can then be obtained by the sample average of u_t^2. In practice, however, u_t's are observed indirectly through the estimates of α and β. Hence a feasible estimator of σ^2 can be obtained by replacing α and β in the definition of u_t by their *OLS* estimators. Namely,

$$\tilde{\sigma}^2 = \frac{\sum_{t=1}^{T} \hat{u}_t^2}{T} = \frac{\sum_{t=1}^{T} \left(y_t - \hat{\alpha} - \hat{\beta} x_t\right)^2}{T},$$

which is the same as the *ML* estimator of σ^2 given by (3). When T is large, this provides a reasonable estimator of σ^2. However, in finite samples a more satisfactory estimator of σ^2 can be obtained by dividing the sum of squares of the residuals by $T - 2$ rather than T. Namely,

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \left(y_t - \hat{\alpha} - \hat{\beta} x_t\right)^2}{T - 2}, \tag{1.34}$$

where '2' is equal to the number of estimated unknown parameters in the simple regression model (here, $\hat{\alpha}$ and $\hat{\beta}$). Unlike $\tilde{\sigma}^2$, the above estimator of σ^2 given by (1.34) is unbiased. Namely $E(\hat{\sigma}^2) = \sigma^2$.

Using the above estimator of σ^2 it is now possible to estimate the variances and covariances of $\hat{\beta}$ given in (1.28). For example we have

$$\widehat{Var}\left(\hat{\beta}\right) = \frac{\hat{\sigma}^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2},$$

and similarly for $\widehat{Var}\left(\hat{\alpha}\right)$ and $\widehat{\text{Cov}}\left(\hat{\alpha}, \hat{\beta}\right)$.

The problem of testing the statistical significance of the estimates and their confidence bands will be addressed in Chapter 3.

1.10 The prediction problem

Suppose T pairs of observations (y_1, x_1), $(y_2, x_2), \ldots (y_T, x_T)$ are available on Y and X and assume that the linear regression of Y on X provides a reasonable model for this T-tuple. The problem of prediction arises when a new observation on X, say x_{T+1}, is considered and it is desired to obtain the 'best' estimate y_{T+1}, the value of Y which corresponds to x_{T+1}. This is called the problem of conditional prediction, namely estimating the value of Y conditional on a *given* value of X. The solution is given by the mathematical expectation of y_{T+1} conditional on the available information, namely $x_1, x_2, \ldots, x_T, x_{T+1}$, and possibly observations on lagged values of Y. In the case of the simple linear regression (1.14) we have

$$E\left(y_{T+1} \left| y_1, y_2, \ldots, y_T; x_1, x_2, \ldots, x_T, x_{T+1}\right.\right) = E\left(y_{T+1} \left| x_{T+1}\right.\right) = \alpha + \beta x_{T+1}.$$

An estimate of this expression gives the *estimate* of the conditional predictor of y_{T+1}. The *OLS* estimate of y_{T+1} is given by

$$\hat{y}_{T+1} = \hat{E}\left(y_{T+1} \left| x_1, x_2, \ldots\right.\right) = \hat{\alpha} + \hat{\beta} x_{T+1}.$$

The variance of the prediction can now be computed as[2]

$$Var\left(\hat{y}_{T+1}\right) = Var\left(\hat{\alpha}\right) + x_{T+1}^2 Var\left(\hat{\beta}\right) + 2x_{T+1}\,\text{Cov}\left(\hat{a}, \hat{\beta}\right).$$

[2] Notice that for the two random variables x and y, and the fixed constants α and β, we have

$$Var\left(\alpha x + \beta y\right) = \alpha^2 Var\left(x\right) + \beta^2 Var\left(y\right) + 2\alpha\beta\,\text{Cov}\left(x, y\right).$$

Now using the results in (1.28) we have:

$$\begin{aligned} Var\left(\hat{y}_{T+1}\right) &= \frac{\sigma^2 \sum_t x_t^2}{T\sum_t (x_t-\bar{x})^2} + \left\{x_{T+1}^2 \frac{\sigma^2}{\sum_t (x_t-\bar{x})^2}\right\} + 2x_{T+1}\left[\frac{-\bar{x}\sigma^2}{\sum_t (x_t-\bar{x})^2}\right] \\ &= \frac{\sigma^2\left[\frac{\sum_t x_t^2}{T} + x_{T+1}^2 - 2\bar{x}x_{T+1}\right]}{\sum_{t=1}^{T}(x_t-\bar{x})^2} \\ &= \frac{\sigma^2}{T}\left[\frac{\sum_t x_t^2 + Tx_{T+1}^2 - 2\left(\sum_t x_t\right)x_{T+1}}{\sum_t (x_t-\bar{x})^2}\right]. \end{aligned}$$

Therefore

$$Var\left(\hat{y}_{T+1}\right) = \frac{\sigma^2}{T\sum_t (x_t-\bar{x})^2}\left[\sum_t (x_t-\bar{x})^2 + T(x_{T+1}-\bar{x})^2\right],$$

or

$$Var\left(\hat{y}_{T+1}\right) = \sigma^2\left[\frac{1}{T} + \frac{(x_{T+1}-\bar{x})^2}{\sum_t (x_t-\bar{x})^2}\right]. \tag{1.35}$$

An estimate of $Var\left(\hat{y}_{T+1}\right)$ is now given by

$$\widehat{Var}\left(\hat{y}_{T+1}\right) = \hat{\sigma}^2\left[\frac{1}{T} + \frac{(x_{T+1}-\bar{x})^2}{\sum_t (x_t-\bar{x})^2}\right]. \tag{1.36}$$

The general theory of prediction under alternative loss functions is discussed in Chapter 17.

1.10.1 Prediction errors and their variance

The error of the conditional forecast of y_{T+1} is defined by

$$\hat{u}_{T+1} = y_{T+1} - \hat{y}_{T+1}.$$

Under the assumption that y_{T+1} is generated according to the simple regression model we have

$$\hat{u}_{T+1} = \alpha + \beta x_{T+1} + u_{T+1} - \hat{\alpha} - \hat{\beta}x_{T+1}.$$

To compute the variance of $\hat{u}_{T+1}$ we first note that both $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of the disturbances over the estimation period (namely $u_1, u_2, \ldots, u_T$) and do not depend on u_{T+1}. Since by assumption u_t's are serially uncorrelated it therefore follows that

$$\begin{aligned} \text{Cov}\left(u_{T+1}, \hat{\alpha} - \alpha\right) &= 0, \\ \text{Cov}\left(u_{T+1}, \hat{\beta} - \beta\right) &= 0. \end{aligned}$$

Hence, conditional on x_{T+1}, u_{T+1} and $\hat{y}_{T+1} = \hat{\alpha} + \hat{\beta}x_{T+1}$ will also be uncorrelated, and

$$Var\left(\hat{u}_{T+1}\right) = Var\left(u_{T+1}\right) + Var\left(\hat{y}_{T+1}\right).$$

Now noting that $Var\left(u_{T+1}\right) = \sigma^2$ and using (1.35) we have

$$Var\left(\hat{u}_{T+1}\right) = \sigma^2\left\{1 + \frac{1}{T} + \frac{(x_{T+1} - \bar{x})^2}{\sum_t (x_t - \bar{x})^2}\right\}. \tag{1.37}$$

This variance can again be estimated by

$$\widehat{Var}\left(\hat{u}_{T+1}\right) = \hat{\sigma}^2\left\{1 + \frac{1}{T} + \frac{(x_{T+1} - \bar{x})^2}{\sum_t (x_t - \bar{x})^2}\right\}.$$

In the case where $\{x_t\}$ has a constant variance, $Var\left(\hat{u}_{T+1}\right)$ converges to σ^2 as $T \to \infty$. The above derivations also clearly show that $Var\left(\hat{u}_{T+1}\right)$ is composed of two parts: one part is due to the inherent uncertainty that surrounds the regression line (i.e., $Var\left(u_t\right) = \sigma^2$), and the other part is due to the sampling variation that is associated with the estimation of the regression parameters, α and β. It is, therefore, natural that as $T \to \infty$, the latter source of variations disappears and we are left with the inherent uncertainty due to the regression, as measured by σ^2.

1.10.2 *Ex ante* predictions

In the case of the linear regression model the *ex ante* prediction of y_{T+1} is obtained without assuming x_{T+1} is known. The prediction is conditional on knowing the past (but not the current) values of x. To obtain *ex ante* prediction of y_{T+1} we therefore also need to predict x_{T+1} conditional on its past values. This requires developing an explicit model for x_t. One popular method of generating *ex ante* forecasts is to assume a univariate time series process for x_ts, and then predict x_{T+1} from information on its lagged values. A simple example of such a time series process is the $AR(1)$ model:

$$x_t = \rho x_{t-1} + \varepsilon_t, \qquad |\rho| < 1,$$

where ε_ts are assumed to have zero means and constant variances. Under this specification the 'optimal' forecast of x_{T+1} (conditional on past values of x's) is given by

$$E\left(x_{T+1} \left| x_1, x_2, \ldots, x_T\right.\right) = \rho x_T,$$

which in turn yields the following *ex ante* forecast of y_{T+1}

$$E\left(y_{T+1} \left| x_1, x_2, \ldots, x_T, y_1, y_2, \ldots, y_T\right.\right) = \alpha + \beta E\left(x_{T+1} \left| x_1, \ldots, x_T\right.\right).$$

An estimate of this forecast is now given by

$$\hat{y}_{T+1} = \hat{E}\left(y_{T+1} \left| x_T\right.\right) = \hat{\alpha} + \hat{\beta}\hat{\rho}x_T,$$

where $\hat{\rho}$ is the *OLS* estimator of ρ, obtained from the regression of x_t on its one-period lagged value. In Chapter 17 we review forecasting within the general context of *ARMA* models, introduced in Chapter 12.

1.11 Exercises

1. Show that the correlation coefficient defined in (1.8) ranges between -1 and 1.
2. In the model $y_t = \alpha + \beta x_t + u_t$ what happens to the *OLS* estimator of β if x_t and/or y_t are standardized by demeaning and scaling by their standard deviations?
3. The following table provides a few key summary statistics for daily rates of change of UK stock index (FTSE) and the GB pound versus US dollar.

Daily UK stock returns and GBP/US$ rate (%)
sample period 2 Jan 1987–16 June 1998

	Stock (FTSE)	**FX (GBP/US$)**
Max	5.69	2.82
Min	−12.11	−3.2861
Mean	0.0396	0.0033
St. dev.	0.8342	0.6200
Skewness	−1.82	−0.27
Kurtosis − 3	26.17	2.55

Using these statistics what do you think are the main differences between these two series and how best these differences are characterized?

4. Consider the following data

Height in centimeters	**Weight in kilograms**
(X)	(Y)
169.6	71.2
166.8	58.2
157.1	56.0
181.1	64.5
158.4	53.0
165.6	52.4
166.7	56.8
156.5	49.2
168.1	55.6
165.3	77.8
$\bar{X} = 165.52$	$\bar{Y} = 59.47$

We obtain

$$S_{XX} = 472.076,$$
$$S_{YY} = 731.961,$$
$$S_{XY} = 274.786.$$

Plot Y against X. Run *OLS* regressions of Y on X and the reverse regression of Y on X. Check that the fitted regression line goes through the means of X and Y.

5. Consider the simple regression model

$$y_t = \alpha + \beta x_t + u_t, \qquad t = 1, 2, \ldots, T$$

where x_t is the explanatory variable and u_t is the unobserved disturbance term.

   (a) Explain briefly what is meant by saying that an estimator, $\hat{\beta}$, of β is:
      i. unbiased
      ii. consistent
      iii. maximum likelihood.
   (b) Under what assumptions is the *OLS* estimator of β:
      i. the best linear unbiased estimator
      ii. the maximum likelihood estimator
   (c) For each of the assumptions you have listed under (b) give an example where the assumption might not hold in economic applications.
   (d) In the model above, why do econometricians make assumptions about the distribution of u_t when testing a hypothesis about the value of β?

6. Consider the following two specifications

$$W_i = a + b \log(E_i) + \varepsilon_i,$$
$$\ln(W_i) = \alpha + \beta \log(E_i) + v_i,$$

where $W_i = P F_i/E_i$, is the share of food expenditure of household i, P is the price of food assumed fixed across all households, $E_i = F_i + NF_i$, with F_i and NF_i are respectively food and non-food expenditures of the household, ε_i and v_i are random errors, a, b, α and β are constant coefficients.

   (a) How do you use these specifications to compute the elasticity of food expenditure relative to the total expenditure?
   (b) Discuss the relative statistical and theoretical merits of these specifications for the analysis of food expenditure.

2 Multiple Regression

2.1 Introduction

This chapter considers the extension of the bivariate regression discussed in Chapter 1 to the case where more than one variable is available to explain/predict y_t, the dependent variable. The topic is known as multiple regression analysis, although only one relationship is in fact considered between y_t and the k explanatory variables, x_{ti}, for $i = 1, 2, \ldots, k$. The problem of multiple regressions where m sets of dependent (or endogenous) variables, y_{tj}, $j = 1, 2, \ldots, m$ are explained in terms of x_{ti}, for $i = 1, 2, \ldots, k$ will be considered in Chapter 19 and is known as multivariate analysis and includes topics such as canonical correlation and factor analysis. This chapter covers standard techniques such as ordinary least squares (*OLS*) and examines the properties of *OLS* estimators under classical assumption, discusses the Gauss–Markov theorem, multiple correlation coefficient, the multicollinearity problem, partitioned regression, introduces regressions that are nonlinear in variables and discusses the interpretation of coefficients.

2.2 The classical normal linear regression model

Consider the general linear regression model

$$y_t = \sum_{j=1}^{k} \beta_j x_{tj} + u_t, \quad \text{for } t = 1, 2, \ldots, T, \tag{2.1}$$

where $x_{t1}, x_{t2}, \ldots, x_{tk}$ are the t^{th} observation on k regressors. If the regression contains an intercept, then one of the k regressors, say the first one x_{t1}, is set equal to unity for all t, namely $x_{t1} = 1$. The parameters $\beta_1, \beta_2, \ldots, \beta_k$ assumed to be fixed (i.e., time invariant) are the regression coefficients, and u_t are the 'disturbances' or the 'errors' of the regression equation. The regression equation can also be written more compactly as

$$y_t = \boldsymbol{\beta}' \mathbf{x}_t + u_t, \quad \text{for } t = 1, 2, \ldots, T, \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)'$ and $\mathbf{x}_t = (x_{t1}, x_{t2}, \ldots, x_{tk})'$. Stacking the equations for all the T observation and using matrix notations, (2.1) or (2.2) can be written as (see Appendix A for an introduction to matrices and matrix operations)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{2.3}$$

where

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tk} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{pmatrix}.$$

The disturbances u_t (or $\mathbf{u}$) satisfy the following assumptions:

Assumption A1: *Zero mean*: the disturbances u_t have zero means

$$E(\mathbf{u}) = \mathbf{0}, \text{ or } E(u_t) = 0, \quad \text{for all} \quad t.$$

Assumption A2: *Homoskedasticity*: the disturbances u_t have constant conditional variances

$$Var(u_t \,|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) = \sigma^2 > 0, \quad \text{for all } t.$$

Assumption A3: *Non-autocorrelated errors*: the disturbances u_t are serially uncorrelated

$$Cov(u_t, u_s \,|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) = 0, \quad \text{for all} \quad t \neq s.$$

Assumption A4: *Orthogonality*: the disturbances u_t and the regressors $x_{t1}, x_{t2}, \ldots, x_{tk}$ are uncorrelated

$$E(u_t \,|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) = 0, \quad \text{for all } t.$$

Assumption A5: *Normality*: the disturbances u_t are normally distributed.

Assumption A2 implies that the variances of u_ts are constant also unconditionally, since,[1]

$$Var\,(u_t) = Var\,[E(u_t \,|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)] + E\,[Var(u_t \,|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)] = \sigma^2,$$

given that, under A4, $E(u_t \,|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) = 0$. The assumption of constant conditional and unconditional error variances is likely to be violated when dealing with cross-sectional regressions, while that of constant conditional error variances is often violated in analysis of financial and macro-economic times series, such as exchange rates, stock returns and interest rates. However, it is possible for errors to be unconditionally constant (time-invariant) but conditionally

[1] See Appendix B, result (B.22).

time varying. Examples include stationary autoregressive conditional heteroskedastic (ARCH) models developed by Engle (1982) and discussed in detail in Chapters 18 and 25.

In time series analysis the critical assumptions are A3 and A4. Assumption A3 is particularly important when the regression equation contains lagged values of the dependent variable, namely $y_{t-1}, y_{t-2}, \ldots$. However, even if lagged values of y_t are not included among the regressors, the breakdown of assumption A3 can lead to misleading inferences, a problem recognized as early as 1920s by Yule (1926), and known in the econometrics time series literature as the spurious regression problem.[2] The orthogonality assumption, A4, allows the empirical analysis of the relationship between y_t and $x_{t1}, x_{t2}, \ldots, x_{tk}$ to be carried out without fully specifying the stochastic processes generating the regressors, also known as 'forcing' variables. We notice that assumption A1 is implied by A4, if a vector of ones is included among the regressors. It is therefore important that an intercept is always included in the regression model, unless it is found to be statistically insignificant.

As they stand, assumptions A2, A3, and A4 require the regressors to be strictly exogenous, in the sense that the first- and second-order moments of the errors, u_t, $t = 1, 2, \ldots, T$, are uncorrelated with the current, past *and* future values of the regressors (see Section 9.3 for a discussion of strict and weak exogeneity, and their impact on the properties of estimators). This assumption is too restrictive for many applications in economics and in effect treats the regressors as given which is more suitable to outcomes of experimental designs rather than economic observations that are based on survey data of transaction prices and quantities. The strict exogeneity assumption also rules out the inclusion of lagged values of y_t amongst the regressors. However, it is possible to relax these assumptions somewhat so that it is only required that the first- and second-order moments of the errors are uncorrelated with current and past values of the regressors, but allowing for the errors to be correlated with the future values of the regressors. In this less restrictive setting, assumptions A2–A4 need to be replaced by the following assumptions:

Assumption A2(i) *Homoskedasticity*: the disturbances u_t have constant conditional variances

$$Var(u_t \,|\mathbf{x}_\ell\,) = \sigma^2 > 0, \quad \text{for all } \ell \leq t.$$

Assumption A3(i) *Non-autocorrelated errors*: the disturbances u_t are serially uncorrelated

$$Cov(u_t, u_s \,|\mathbf{x}_\ell\,) = 0, \quad \text{for all} \quad t \neq s \text{ and } \ell \leq \min(t, s).$$

Assumption A4(i) *Orthogonality*: the disturbances u_t and the regressors $x_{t1}, x_{t2}, \ldots, x_{tk}$ are uncorrelated

$$E(u_t \,|\mathbf{x}_\ell\,) = 0, \quad \text{for all } \ell \leq t.$$

Under these assumptions the regressors are said to be *weakly exogenous*, and allow for lagged values of y_t to be included in $\mathbf{x}_t$.

Adding assumption A5 to the classical model yields the classical linear normal regression model. This model can also be derived using the *joint* distribution of y_t, $\mathbf{x}_t$, and by assuming

[2] Champernowne (1960) and Granger and Newbold (1974) provide Monte Carlo evidence on the spurious regression problem, and Phillips (1986) establishes a number of theoretical results.

that this distribution is a multivariate *normal* with constant means, variances and covariances. In this setting, the regression of y_t on $\mathbf{x}_t$, defined as the mathematical expectation of y_t conditional on the realized values of the regressors, will be linear in the regressors. The linearity of the regression equation follows from the joint normality assumption and need not hold if this assumption is relaxed. To be more precise suppose that

$$\begin{pmatrix} y_t \\ \mathbf{x}_t \end{pmatrix} \backsim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.4}$$

where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \text{ and } \Sigma = \begin{pmatrix} \sigma_{yy} & \boldsymbol{\sigma}_{yx} \\ \boldsymbol{\sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix}.$$

Then using known results from theory of multivariate normal distributions (see Appendix B for a summary and references) we have

$$\begin{aligned} E(y_t | \mathbf{x}_t) &= \mu_y + \boldsymbol{\sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_x), \\ Var(y_t | \mathbf{x}_t) &= \sigma_{yy} - \boldsymbol{\sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\sigma_{xy}. \end{aligned}$$

Under this setting, assuming that (2.2) includes an intercept, the regression coefficients $\boldsymbol{\beta}$ will be given by $(\mu_y - \boldsymbol{\sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\mu}_x, \boldsymbol{\sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1})'$. It is also easily seen that the regression errors associated with (2.4) are given by

$$u_t = y_t - (\mu_y - \boldsymbol{\sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\mu}_x) - \boldsymbol{\sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{x}_t,$$

and, by construction, satisfy the classical assumptions. But note that no dynamic effects are allowed in the distribution of $(y_t, \mathbf{x}_t')'$.

Both of the above interpretations of the classical normal regression model have been used in the literature (see, e.g., Spanos (1989)). We remark that the normality assumption A5 may be important in small samples, but is not generally required when the sample under consideration is large enough.

All the various departures from the classical normal regression model mentioned here will be analysed in Chapters 3 to 6.

2.3 The method of ordinary least squares in multiple regression

The criterion function in this general case will be

$$Q(\beta_1, \beta_2, \ldots, \beta_k) = \sum_{t=1}^{T} \left(y_t - \sum_{j=1}^{k} \beta_j x_{tj} \right)^2. \tag{2.5}$$

The necessary conditions for the minimization of $Q\left(\beta_1, \beta_2, \ldots, \beta_k\right)$ are given by

$$\frac{\partial Q\left(\beta_1, \beta_2, \ldots, \beta_k\right)}{\partial \beta_s} = -2\sum_{t=1}^{T} x_{ts}\left(y_t - \sum_{j=1}^{k}\hat{\beta}_j x_{tj}\right) = 0, \qquad s = 1, 2, \ldots, k, \tag{2.6}$$

where $\hat{\beta}_j$ is the *OLS* estimator of β_j. The k equations in (2.6) are known as the 'normal' equations. Denoting the residuals by $\hat{u}_t = y_t - \sum_j \hat{\beta}_j x_{tj}$, the normal equations can be written as $\sum_{t=1}^{T} x_{ts}\hat{u}_t = 0$, for $s = 1, 2, \ldots, k$, or, in expanded form

$$\begin{aligned}\sum_{t=1}^{T} x_{ts}y_t &= \sum_{t=1}^{T}\sum_{j=1}^{k}\hat{\beta}_j x_{tj}x_{ts} \\ &= \sum_{j=1}^{k}\hat{\beta}_j\left(\sum_{t=1}^{T} x_{tj}x_{ts}\right).\end{aligned}$$

Without the use of matrix notations, the study of the properties of multiple regression would be extremely tedious. In matrix form, the criterion function (2.5) to be minimized is

$$Q\left(\boldsymbol{\beta}\right) = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right), \tag{2.7}$$

and the first-order conditions become

$$\frac{\partial Q\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{0},$$

which yield the normal equations,

$$\left(\mathbf{X}'\mathbf{X}\right)\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}.$$

Suppose now that $\mathbf{X}'\mathbf{X}$ is of full rank, that is Rank $\left(\mathbf{X}'\mathbf{X}\right) = k$, [or Rank$(\mathbf{X}) = k$] a necessary condition for this is that $k \leq T$. There should be at least as many observations as there are unknown coefficients. Then

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}. \tag{2.8}$$

In the case where Rank$(\mathbf{X}) = r < k$, $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-}\mathbf{X}'\mathbf{y}$, where $\left(\mathbf{X}'\mathbf{X}\right)^{-}$ represents the generalized inverse of $\mathbf{X}'\mathbf{X}$. In this case only r linear combinations of the regression coefficients are uniquely determined.

2.4 The maximum likelihood approach

Under the normality assumption A5, the *OLS* estimator can be derived by maximizing the likelihood function associated to model (2.2) (or (2.3)). Let $\boldsymbol{\theta} = \left(\boldsymbol{\beta}', \sigma^2\right)'$, then the likelihood of

a sample of T independent, identically and normally distributed disturbances is[3]

$$L_T(\boldsymbol{\theta}) = \left(2\pi\sigma^2\right)^{-T/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^{T} u_t^2\right)$$
$$= \left(2\pi\sigma^2\right)^{-T/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t - \boldsymbol{\beta}'\mathbf{x}_t\right)^2\right].$$

Adopting the matrix notation,

$$L_T(\boldsymbol{\theta}) = \left(2\pi\sigma^2\right)^{-T/2} \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right]. \tag{2.9}$$

Taking logs, we obtain the log-likelihood function for the classical linear regression model

$$\ell_T(\boldsymbol{\theta}) = \log L_T(\boldsymbol{\theta}) = -\frac{T}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\left(y_t - \boldsymbol{\beta}'\mathbf{x}_t\right)^2. \tag{2.10}$$

The necessary conditions for maximizing (2.10) are

$$\begin{pmatrix} \frac{\partial \ell_T(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \\ \frac{\partial \ell_T(\boldsymbol{\theta})}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2}\mathbf{X}'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \\ -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ 0 \end{pmatrix}.$$

The values that satisfy these equations are

$$\widetilde{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}, \text{ and } \widetilde{\sigma}^2 = \frac{\sum_t \widetilde{u}_t^2}{T} = \frac{\widetilde{\mathbf{u}}'\widetilde{\mathbf{u}}}{T},$$

where $\widetilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}$. Notice that the estimator for the slope coefficients is identical to the *OLS* estimator (2.8), while the variance estimator differs from (2.14) by the divisor of T instead of $T - k$. Clearly, the *OLS* estimator inherits all the asymptotic properties of the *ML* estimator. We refer to Chapter 9 for a review of the theory underlying the maximum likelihood approach, and to Chapter 19 for an extension of the above results to the case of multivariate regression.

The likelihood approach also forms the basis of the Bayesian inference where the likelihood is combined with prior distributions on the unknown parameters to obtain posterior probability distributions which is then used for estimation and inference: see Section C.6 in Appendix C.

[3] See also Section 1.8 where the likelihood approach is introduced for the analysis of bivariate regression models.

2.5 Properties of *OLS* residuals

The residual vector is given by

$$\begin{aligned}\hat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &= \left[\mathbf{I}_T - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{y} \\ &= \mathbf{M}\mathbf{y},\end{aligned}$$

where $\mathbf{I}_T$ is an identity matrix of order T, $\mathbf{M} = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, with the property $\mathbf{M}^2 = \mathbf{M}$, which makes $\mathbf{M}$ to be an *idempotent* matrix. Also $\mathbf{M} = \mathbf{I}_T - \mathbf{P}$, where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is called the *projection matrix* of the regression (2.3). Note that

$$\begin{aligned}\mathbf{M}\mathbf{X} &= \left[\mathbf{I}_T - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\mathbf{X} \\ &= \mathbf{X} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X} \\ &= \mathbf{X} - \mathbf{X} = \mathbf{0}.\end{aligned}$$

Therefore

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{X}'\mathbf{M}\mathbf{y} = \mathbf{0}, \tag{2.11}$$

or $\sum_{t=1}^{T} x_{ts}\hat{u}_t = 0$, for $s = 1, 2, \ldots, k$ which are the normal equations of the regression problem. Therefore, the regressors are by construction 'orthogonal' to the vector of *OLS* residuals.

In the case where the regression equation contains an intercept term (i.e., when one of the x_{tj}'s is equal to 1 for all t) we also have

$$\sum_{t=1}^{T}\hat{u}_t = 0 = T\left(\bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 - \ldots - \hat{\beta}_k\bar{x}_k\right) = 0,$$

where $\bar{x}_j$ stands for the sample mean of the j^{th} regressor, x_{tj}. This result follows directly from the normal equations $\sum_{t=1}^{T} x_{ts}\hat{u}_t = 0$, by choosing x_{ts} to be the intercept term, namely setting $x_{ts} = 1$ in $\sum_{t=1}^{T} x_{ts}\hat{u}_t = 0$.

To summarize, the *OLS* residual vector, $\hat{\mathbf{u}}$, has the following properties:

(i) By construction all the regressors are orthogonal to the residual vector, that is, $\mathbf{X}'\hat{\mathbf{u}} = 0$.

(ii) When the regression equation contains an intercept term, the residuals, $\hat{u}_t$, have mean zero exactly, i.e. $\sum_{t=1}^{T}\hat{u}_t = 0$. This result also implies that the regression plane goes through the sample mean of $\mathbf{y}$ and the sample means of all the regressors.

(iii) Even if u_t are homoskedastic and serially uncorrelated, the *OLS* residuals, $\hat{u}_t$, will be *heteroskedastic* and *autocorrelated* in small samples.

Result (iii) follows by noting that

$$\hat{\mathbf{u}} = \mathbf{M}\mathbf{y} = \mathbf{M}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right) = \mathbf{M}\mathbf{u},$$

and

$$E\left(\hat{\mathbf{u}}\hat{\mathbf{u}}'\right) = E\left(\mathbf{M}\mathbf{u}\mathbf{u}'\mathbf{M}'\right) = \mathbf{M}E\left(\mathbf{u}\mathbf{u}'\right)\mathbf{M}'.$$

But under the classical assumptions $E\left(\mathbf{u}\mathbf{u}'\right) = \sigma^2\mathbf{I}_T$. Hence

$$E\left(\hat{\mathbf{u}}\hat{\mathbf{u}}'\right) = \mathbf{M}\left(\sigma^2\mathbf{I}_T\right)\mathbf{M}' = \sigma^2\mathbf{M}\mathbf{M}' = \sigma^2\mathbf{M},$$

which is different from an identity matrix and establishes that $\hat{u}_t$ and $\hat{u}_{t'}$ $\left(t \neq t'\right)$ are neither uncorrelated nor homoskedastic. These properties of *OLS* residuals lie at the core of some of the difficulties encountered in practice in developing tests of the classical assumptions based on *OLS* residuals, that perform well in small samples. Fortunately, the serial correlation and heteroskedasticity properties of *OLS* residuals tend to disappear in 'large enough' samples.

2.6 Covariance matrix of $\hat{\boldsymbol{\beta}}$

The covariance matrix of $\hat{\boldsymbol{\beta}}$ is defined as

$$\begin{aligned} Var\left(\hat{\boldsymbol{\beta}}\right) &= E\left\{\left[\hat{\boldsymbol{\beta}} - E\left(\hat{\boldsymbol{\beta}}\right)\right]\left[\hat{\boldsymbol{\beta}} - E\left(\hat{\boldsymbol{\beta}}\right)\right]'\right\} \\ &= \begin{pmatrix} Var\left(\hat{\beta}_1\right) & \text{Cov}\left(\hat{\beta}_1, \hat{\beta}_2\right) & \cdots & \text{Cov}\left(\hat{\beta}_1, \hat{\beta}_k\right) \\ \text{Cov}\left(\hat{\beta}_2, \hat{\beta}_1\right) & Var\left(\hat{\beta}_2\right) & \cdots & \text{Cov}\left(\hat{\beta}_2, \hat{\beta}_k\right) \\ \vdots & & & \\ \text{Cov}\left(\hat{\beta}_k, \hat{\beta}_1\right) & \text{Cov}\left(\hat{\beta}_k, \hat{\beta}_2\right) & \cdots & Var\left(\hat{\beta}_k\right) \end{pmatrix}. \end{aligned} \tag{2.12}$$

The diagonal elements of the matrix $Var\left(\hat{\boldsymbol{\beta}}\right)$ are the variances of the *OLS* estimators, $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k\right)'$, and the off-diagonal elements are the covariances.

To obtain the formula for $Var\left(\hat{\boldsymbol{\beta}}\right)$ we first note that

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

But $E(\mathbf{X}'\mathbf{u}\,|\mathbf{X}\,) = \mathbf{X}'E(\mathbf{u}\,|\mathbf{X}\,)$ and, under assumption A4, $E(\mathbf{u}\,|\mathbf{X}\,) = \mathbf{0}$, and hence $E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$, namely that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$. Also

$$\hat{\boldsymbol{\beta}} - E\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}.$$

Therefore

$$Var\left(\hat{\boldsymbol{\beta}}\right) = E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right].$$

Again under assumption A4

$$E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\mathbf{X}\right]=\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E\left(\mathbf{u}\mathbf{u}'|\mathbf{X}\right)\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

and under assumptions A2 and A3, $E\left(\mathbf{u}\mathbf{u}'|\mathbf{X}\right)=\sigma^2\mathbf{I}_T$. Therefore,

$$E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}\mathbf{u}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}|\mathbf{X}\right]=\sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

and hence

$$Var\left(\hat{\boldsymbol{\beta}}\right)=\sigma^2E\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]. \tag{2.13}$$

For given values of $\mathbf{X}$ an estimator of $Var\left(\hat{\boldsymbol{\beta}}\right)$ is

$$\widehat{Var}\left(\hat{\boldsymbol{\beta}}\right)=\hat{\sigma}^2\left(\mathbf{X}'\mathbf{X}\right)^{-1},$$

where $\hat{\sigma}^2$ is

$$\hat{\sigma}^2=\frac{\sum_t\hat{u}_t^2}{T-k}=\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k}, \tag{2.14}$$

with k being the number of regressors, including the intercept term. As in the case of the simple regression model, $\hat{\sigma}^2$ is an unbiased estimator of σ^2, namely $E(\hat{\sigma}^2)=\sigma^2$. Unbiasedness of $\hat{\sigma}^2$ is easily established by noting that $\hat{\mathbf{u}}=\mathbf{M}\mathbf{u}$ and hence

$$\begin{aligned}E\left(\hat{\sigma}^2\right)&=\left(\frac{1}{T-k}\right)E\left(\mathbf{u}'\mathbf{M}\mathbf{u}\right)\\&=\left(\frac{1}{T-k}\right)E\left[Tr\left(\mathbf{u}'\mathbf{M}\mathbf{u}\right)\right]=\left(\frac{1}{T-k}\right)E\left[Tr\left(\mathbf{u}\mathbf{M}\mathbf{u}'\right)\right]\\&=\left(\frac{1}{T-k}\right)Tr\left[\mathbf{M}E\left(\mathbf{u}\mathbf{u}'\right)\right]=\left(\frac{1}{T-k}\right)Tr\left(\mathbf{M}\sigma^2\right),\end{aligned}$$

Noting that

$$\begin{aligned}Tr\left(\mathbf{M}\right)&=Tr\left[\mathbf{I}_T-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]\\&=Tr\left(\mathbf{I}_T\right)-Tr\left[\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]=T-k,\end{aligned}$$

it follows that

$$E\left(\hat{\sigma}^2\right)=\frac{\sigma^2Tr\left(\mathbf{M}\right)}{T-k}=\sigma^2.$$

The estimator of Cov $\left(\hat{\beta}_j, \hat{\beta}_s\right)$ is given by the $(j,s)^{th}$ element of matrix $\hat{\sigma}^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$.

Example 1 *Consider the three variable regression model*

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + u_t, \qquad t = 1, 2, \ldots, T, \tag{2.15}$$

where we have set the first variable, x_{t1}, equal to unity to allow for an intercept in the regression. To simplify the derivations we work with variables in terms of their deviations from their respective sample means. Summing the equation (2.15) over t and dividing by the sample size, T, yields:

$$\bar{y} = \beta_1 + \beta_2 \bar{x}_2 + \beta_3 \bar{x}_3 + \bar{u}, \tag{2.16}$$

where $\bar{y} = \sum_t y_t/T$, $\bar{x}_2 = \sum_t x_{t2}/T$, $\bar{x}_3 = \sum_t x_{t3}/T$, $\bar{u} = \sum_t u_t/T$ are the sample means. Subtracting (2.16) from (2.15) we obtain

$$y_t - \bar{y} = \beta_2 \left(x_{t2} - \bar{x}_2\right) + \beta_3 \left(x_{t3} - \bar{x}_3\right) + \left(u_t - \bar{u}\right).$$

The OLS estimators of β_2 and β_3 are now given by (using (2.8))

$$\begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} S_{22} & S_{23} \\ S_{23} & S_{33} \end{pmatrix}^{-1} \begin{pmatrix} S_{2y} \\ S_{3y} \end{pmatrix},$$

where

$$\begin{aligned} S_{js} = \textstyle\sum_t \left(x_{tj} - \bar{x}_j\right)\left(x_{ts} - \bar{x}_s\right) = \sum_t \left(x_{tj} - \bar{x}_j\right) x_{ts}, \qquad & j, s = 2, 3, \\ S_{jy} = \textstyle\sum_t \left(x_{tj} - \bar{x}_j\right) y_t, \qquad & j = 2, 3, \end{aligned}$$

$$\begin{pmatrix} S_{22} & S_{23} \\ S_{23} & S_{33} \end{pmatrix}^{-1} = \frac{1}{S_{22}S_{33} - S_{23}^2} \begin{bmatrix} S_{33} & -S_{23} \\ -S_{23} & S_{22} \end{bmatrix}.$$

Hence

$$\hat{\beta}_2 = \frac{S_{33}S_{2y} - S_{23}S_{3y}}{S_{22}S_{33} - S_{23}^2}, \tag{2.17}$$

$$\hat{\beta}_3 = \frac{S_{22}S_{3y} - S_{23}S_{2y}}{S_{22}S_{33} - S_{23}^2}. \tag{2.18}$$

The estimator of β_1, the intercept term, can now be obtained recalling that the regression plane goes through the sample means when the equation has an intercept term. Namely

$$\bar{y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{x}_2 + \hat{\beta}_3 \bar{x}_3,$$

and hence

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2 \bar{x}_2 - \hat{\beta}_3 \bar{x}_3. \tag{2.19}$$

The estimates of the variances and the covariance of $\hat{\beta}_2$ and $\hat{\beta}_3$ are given by [using (2.12) and (2.13)]

$$\widehat{Cov}\begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \hat{\sigma}^2 \begin{pmatrix} S_{22} & S_{23} \\ S_{23} & S_{33} \end{pmatrix}^{-1},$$

or

$$\widehat{Var}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}^2 S_{33}}{S_{22}S_{33} - S_{23}^2}, \tag{2.20}$$

$$\widehat{Var}\left(\hat{\beta}_3\right) = \frac{\hat{\sigma}^2 S_{22}}{S_{22}S_{33} - S_{23}^2}, \tag{2.21}$$

and

$$\widehat{Cov}\left(\hat{\beta}_2, \hat{\beta}_3\right) = -\frac{\hat{\sigma}^2 S_{23}}{S_{22}S_{33} - S_{23}^2}. \tag{2.22}$$

Finally,

$$\hat{\sigma}^2 = \frac{\sum_t \hat{u}_t^2}{T-3} = \frac{\sum_t \left(y_t - \hat{\beta}_1 - \hat{\beta}_2 x_{t2} - \hat{\beta}_3 x_{t3}\right)^2}{T-3}. \tag{2.23}$$

Notice that the denominator of $\hat{\sigma}^2$ is $T-3$, as we have estimated three coefficients, namely the intercept term, β_1, and the two regression coefficients, β_2 and β_3.

2.7 The Gauss–Markov theorem

The Gauss–Markov theorem states that under the classical assumptions A1–A4 the *OLS* estimator (2.8) has the least variance in the class of all linear unbiased estimators of $\boldsymbol{\beta}$, namely it is the best linear unbiased estimator (BLUE). More formally, let $\boldsymbol{\beta}^*$ be an alternative linear unbiased estimator of $\boldsymbol{\beta}$ defined by

$$\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}} + \mathbf{C}'\mathbf{y}, \tag{2.24}$$

where $\mathbf{C}$ is a $k \times T$ matrix with elements possibly depending on $\mathbf{X}$, but not on $\mathbf{y}$. It is clear that $\boldsymbol{\beta}^*$ is a linear estimator. Also since $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$, for $\boldsymbol{\beta}^*$ to be an unbiased estimator we need

$$E(\boldsymbol{\beta}^*) = E\left(\hat{\boldsymbol{\beta}}\right) + \mathbf{C}'E(\mathbf{y}) = \boldsymbol{\beta} + \mathbf{C}'E(\mathbf{y}) = \boldsymbol{\beta},$$

or that $\mathbf{C}'E(\mathbf{y}) = \mathbf{0}$, which in turn implies that

$$\mathbf{C}'E(\mathbf{y}) = \mathbf{C}'(\mathbf{X}\boldsymbol{\beta} + E(\mathbf{u})) = \mathbf{C}'\mathbf{X}\boldsymbol{\beta} = 0, \tag{2.25}$$

for all values of $\boldsymbol{\beta}$.

To prove the Gauss–Markov we need to show that subject to the unbiasedness condition (2.25), $Var(\hat{\boldsymbol{\beta}}) \leq Var(\boldsymbol{\beta}_*)$, in the sense that $Var(\boldsymbol{\beta}_*) - Var(\hat{\boldsymbol{\beta}})$ is a semi-positive definite matrix. Using (2.3) and (2.8) in (2.24), we have

$$\begin{aligned}\boldsymbol{\beta}^* &= \left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' + \mathbf{C}'\right]\mathbf{y} \\ &= \left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' + \mathbf{C}'\right](\mathbf{X}\boldsymbol{\beta} + \mathbf{u}),\end{aligned}$$

or

$$\boldsymbol{\beta}^* - \boldsymbol{\beta} = \mathbf{C}'\mathbf{X}\boldsymbol{\beta} + \left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' + \mathbf{C}'\right]\mathbf{u}.$$

But using (2.25), $\mathbf{C}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ and

$$\boldsymbol{\beta}^* - \boldsymbol{\beta} = \left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' + \mathbf{C}'\right]\mathbf{u}.$$

Hence (for a given set of observations, $\mathbf{X}$)

$$\begin{aligned}Var(\boldsymbol{\beta}^*) &= E\left[(\boldsymbol{\beta}^* - \boldsymbol{\beta})(\boldsymbol{\beta}^* - \boldsymbol{\beta})'\right] \\ &= \sigma^2\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1} + \mathbf{C}'\mathbf{C} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{C} + \mathbf{C}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right].\end{aligned}$$

However, since $\mathbf{C}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$ for all parameter values, $\boldsymbol{\beta}$, then we should also have $\mathbf{C}'\mathbf{X} = 0$, and

$$Var(\boldsymbol{\beta}^*) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1} + \sigma^2\left(\mathbf{C}'\mathbf{C}\right).$$

Therefore

$$Var(\boldsymbol{\beta}^*) - Var\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{C}'\mathbf{C}\right),$$

which is a semi-positive definite matrix.

The Gauss–Markov theorem readily extends to the *OLS* estimator of any linear combination of the parameters, $\boldsymbol{\beta}$. Consider, for example, the linear combination

$$\delta = \boldsymbol{\lambda}'\boldsymbol{\beta},$$

where λ is a $k \times 1$ vector of fixed coefficients. Denote the *OLS* estimator of δ by $\hat{\delta}$, and the alternative linear unbiased estimator by δ^*. We have

$$\hat{\delta} = \boldsymbol{\lambda}'\hat{\boldsymbol{\beta}}, \qquad \delta^* = \boldsymbol{\lambda}'\boldsymbol{\beta}^*,$$

and

$$\begin{aligned} Var\left(\delta^*\right) - Var\left(\hat{\delta}\right) &= \boldsymbol{\lambda}' Var\left(\boldsymbol{\beta}^*\right)\boldsymbol{\lambda} - \boldsymbol{\lambda}' Var\left(\hat{\boldsymbol{\beta}}\right)\boldsymbol{\lambda} \\ &= \boldsymbol{\lambda}'\left[Var\left(\boldsymbol{\beta}^*\right) - Var\left(\hat{\boldsymbol{\beta}}\right)\right]\boldsymbol{\lambda}. \end{aligned}$$

But we have already shown that $Var\left(\boldsymbol{\beta}^*\right) - Var(\hat{\boldsymbol{\beta}})$ is a semi-positive definite matrix. Therefore,

$$Var\left(\delta^*\right) - Var\left(\hat{\delta}\right) \geq 0.$$

A number of other interesting results also follow from this last inequality. Setting $\boldsymbol{\lambda}' = (1, 0, \ldots, 0)$, for example, gives

$$\delta = \boldsymbol{\lambda}'\boldsymbol{\beta} = \beta_1,$$

and establishes that

$$Var\left(\beta_1^*\right) - Var\left(\hat{\beta}_1\right) \geq \mathbf{0}.$$

Similarly, $Var(\beta_j^*) - Var(\hat{\beta}_j) \geq 0$, for $j = 1, 2, \ldots, k$.

It is important to bear in mind that the Gauss–Markov theorem does not apply if the regressors are weakly exogenous even if all the other assumptions of the classical model are satisfied.

2.8 Mean square error of an estimator and the bias-variance trade-off

The Gauss–Markov theorem states that, under the classical assumptions, it is not possible to find linear unbiased estimators of regression coefficients which have smaller variances than the *OLS* estimator, (2.8). However, as shown by James and Stein (1961), it is possible to find other estimators that are biased but have a lower variance than the *OLS* estimator. The trade-off between bias and variance can be formalized if the alternative estimators are compared by their mean square error defined as

$$MSE(\widetilde{\boldsymbol{\beta}}) = E\left[(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\right],$$

where $\widetilde{\boldsymbol{\beta}}$ denotes an alternative estimator to the *OLS* estimator, $\hat{\boldsymbol{\beta}}$, and $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$. To see the bias-variance trade-off we first note that

$$E\left[(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)(\widetilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\right] = E\left\{\left[(\widetilde{\boldsymbol{\beta}}-E(\widetilde{\boldsymbol{\beta}}))-(\boldsymbol{\beta}_0-E(\widetilde{\boldsymbol{\beta}}))\right]\left[(\widetilde{\boldsymbol{\beta}}-E(\widetilde{\boldsymbol{\beta}}))-(\boldsymbol{\beta}_0-E(\widetilde{\boldsymbol{\beta}}))\right]'\right\}$$
$$= E\left\{\left[\widetilde{\boldsymbol{\beta}}-E(\widetilde{\boldsymbol{\beta}})\right]\left[\widetilde{\boldsymbol{\beta}}-E(\widetilde{\boldsymbol{\beta}})\right]'\right\} + E\left\{\left[\boldsymbol{\beta}_0-E(\widetilde{\boldsymbol{\beta}})\right]\left[\boldsymbol{\beta}_0-E(\widetilde{\boldsymbol{\beta}})\right]'\right\}$$
$$- E\left\{\left[\widetilde{\boldsymbol{\beta}}-E(\widetilde{\boldsymbol{\beta}})\right]\left[\boldsymbol{\beta}_0-E(\widetilde{\boldsymbol{\beta}})\right]'\right\} - E\left\{\left[\boldsymbol{\beta}_0-E(\widetilde{\boldsymbol{\beta}})\right]\left[\widetilde{\boldsymbol{\beta}}-E(\widetilde{\boldsymbol{\beta}})\right]'\right\}.$$

But $\boldsymbol{\beta}_0 - E(\widetilde{\boldsymbol{\beta}})$ is a constant (i.e., non-stochastic), and can be taken outside of the expectations operator. Also

$$E\left\{\left[\widetilde{\boldsymbol{\beta}}-E(\widetilde{\boldsymbol{\beta}})\right]\left[\widetilde{\boldsymbol{\beta}}-E(\widetilde{\boldsymbol{\beta}})\right]'\right\} = Var\left(\widetilde{\boldsymbol{\beta}}\right),$$

and by construction

$$E\left[\widetilde{\boldsymbol{\beta}}-E(\widetilde{\boldsymbol{\beta}})\right] = \mathbf{0}.$$

Hence

$$MSE(\widetilde{\boldsymbol{\beta}}) = Var\left(\widetilde{\boldsymbol{\beta}}\right) + \left[\boldsymbol{\beta}_0 - E(\widetilde{\boldsymbol{\beta}})\right]\left[\boldsymbol{\beta}_0 - E(\widetilde{\boldsymbol{\beta}})\right]'.$$

Namely, the $MSE(\widetilde{\boldsymbol{\beta}})$ can be decomposed into a variance term plus the square of the bias. In principle it is clearly possible to find an estimator for $\boldsymbol{\beta}$ with lower variance at the expense of some bias, leading to a reduction in the overall MSE. This result has been used by James and Stein (1961) to propose a biased estimator for $\boldsymbol{\beta}$ such that its MSE is smaller than the MSE of $\hat{\boldsymbol{\beta}}$. Specifically, they considered the estimator

$$\widetilde{\beta}_j = \left[1 - \frac{(k-2)\,\sigma^2}{\hat{\boldsymbol{\beta}}'\,(\mathbf{X}\mathbf{X}')\,\hat{\boldsymbol{\beta}}}\right]\hat{\beta}_j, \quad j = 1, 2, \ldots, k,$$

obtained by minimizing the overall MSE of $\widetilde{\boldsymbol{\beta}}$. James and Stein proved that this estimator, by shrinking the *OLS* estimator towards zero, has a MSE smaller than the MSE of *OLS* estimator when $k > 2$. For further details see, for example, Draper and van Nostrand (1979) and Gruber (1998).

2.9 Distribution of the *OLS* estimator

Under the classical normal assumptions A1–A5, for a given realization of the regressors, $\mathbf{X}$, the *OLS* estimator, $\hat{\boldsymbol{\beta}}$, is a linear function of u_t, for $t = 1, 2, \ldots, T$, and hence is also normally distributed. More specifically, using (2.8), note that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u},$$

and since under assumptions A1–A5, $\mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T)$, then recalling that $\mathbf{X}'\mathbf{X}$ is a positive definite matrix, we have

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim N[\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}].$$

Equivalently,

$$(\mathbf{X}'\mathbf{X})^{1/2}\left(\frac{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}}{\sigma}\right) \sim N(\mathbf{0}, \mathbf{I}_k),$$

and

$$\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \sim \chi^2_k, \tag{2.26}$$

where χ^2_k stands for the central chi-square distribution with k degrees of freedom. The above result also follows unconditionally.

Consider now the distribution of $\hat{\sigma}^2$, the unbiased estimator of σ^2, given by (2.14). We note that $\hat{\mathbf{u}} = \mathbf{M}\mathbf{u}$, where $\mathbf{M} = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is an idempotent matrix with rank $T-k$. Then the singular value decomposition of $\mathbf{M}$ is given by $\mathbf{G}\mathbf{M}\mathbf{G}' = \boldsymbol{\Lambda}$, where $\mathbf{G}$ is an orthonormal matrix such that $\mathbf{G}\mathbf{G}' = \mathbf{I}_T$, and

$$\boldsymbol{\Lambda} = \begin{pmatrix} \mathbf{I}_{T-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Hence

$$\frac{(T-k)\,\hat{\sigma}^2}{\sigma^2} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\sigma^2} = \frac{\mathbf{u}'\mathbf{M}\mathbf{u}}{\sigma^2} = \boldsymbol{\xi}'\boldsymbol{\Lambda}\boldsymbol{\xi},$$

where $\boldsymbol{\xi} = \sigma^{-1}\mathbf{G}\mathbf{u} \sim N(\mathbf{0}, \mathbf{I}_T)$. Partition $\boldsymbol{\xi}$ conformable to $\boldsymbol{\Lambda}$, and note that

$$\frac{(T-k)\,\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^{T-k} \xi_i^2,$$

where ξ_i are independently and identically distributed as $N(0, 1)$. Thus

$$\frac{(T-k)\,\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{T-k}. \tag{2.27}$$

Finally, using (2.26) and the above result, we have

$$\frac{T-k}{k}\frac{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{\frac{(T-k)\hat{\sigma}^2}{\sigma^2}} = \frac{\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'(\mathbf{X}'\mathbf{X})\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)}{k\hat{\sigma}^2} \sim F(k, T-k),$$

where $F(k, T-k)$ stands for the central F-distribution with k and $T-k$ degrees of freedom. This result follows immediately from the definition of F-distribution, which is given by the ratio

of two independent chi-squared variates corrected for their respective degrees of freedom (see Appendix B). In the present application, the two chi-squared distributions are

$$\frac{(T-k)\,\hat{\sigma}^2}{\sigma^2} = \frac{\mathbf{u}'\mathbf{M}\mathbf{u}}{\sigma^2} \sim \chi^2_{T-k},$$

and

$$\begin{aligned}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &= \mathbf{u}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \frac{\mathbf{u}'(\mathbf{I}_T-\mathbf{M})\mathbf{u}}{\sigma^2} \sim \chi^2_k.\end{aligned}$$

The independence of $\mathbf{u}'\mathbf{M}\mathbf{u}$ and $\mathbf{u}'(\mathbf{I}_T-\mathbf{M})\mathbf{u}$ follows from the fact that $(\mathbf{I}_T-\mathbf{M})\mathbf{M} = \mathbf{M}-\mathbf{M}^2 = \mathbf{M}-\mathbf{M} = \mathbf{0}$.

The above results can be readily adapted for deriving the distribution of linear subsets of $\hat{\boldsymbol{\beta}}$. Suppose we are interested in the distribution of $\mathbf{R}\hat{\boldsymbol{\beta}}$, where $\mathbf{R}$ is an $r \times k$ matrix of fixed constants with rank $r \leq k$. Then

$$\begin{aligned}&\frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{R}\boldsymbol{\beta}\right)'\left[\mathbf{R}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{R}'\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{R}\boldsymbol{\beta}\right)}{r\hat{\sigma}^2} \\ &= \frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{R}\boldsymbol{\beta}\right)'\left[\mathbf{R}\widehat{Var}(\hat{\boldsymbol{\beta}})\mathbf{R}'\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{R}\boldsymbol{\beta}\right)}{r} \sim F(r, T-k).\end{aligned} \tag{2.28}$$

In the case where $r = 1$, the F-test reduces to a t-test. For example, by setting $\mathbf{R} = (1, 0, \ldots, 0)$, the above result implies

$$\frac{(\hat{\beta}_1-\beta_1)^2}{\widehat{Var}(\hat{\beta}_1)} \sim F(1, T-k),$$

which in turn yields the familiar t-test statistic, given by $\left(\hat{\beta}_1-\beta_1\right)/\sqrt{\widehat{Var}(\hat{\beta}_1)} \sim t_{T-k}$.

2.10 The multiple correlation coefficient

By analogy to the case of the simple regression model, the strength of the fit of a multiple regression equation is measured *via* the multiple correlation coefficient, R, defined by the proportion of the total variation of y explained by the regression equation:

$$R^2 = \frac{\sum_t \left(\hat{y}_t-\bar{y}\right)^2}{\sum_t \left(y_t-\bar{y}\right)^2}. \tag{2.29}$$

As in the case of the simple regression equation, the total variation of y, measured by $S_{yy} = \sum_t (y_t - \bar{y})^2$, can be decomposed into that explained by the regression equation, $\sum_t (\hat{y}_t - \bar{y})^2$, and the rest:[4]

$$\sum_t (y_t - \bar{y})^2 = \sum_t (\hat{y}_t - \bar{y})^2 + \sum_t (y_t - \hat{y}_t)^2 .$$

Hence, R^2 can also be written as

$$R^2 = \frac{\sum_t (y_t - \bar{y})^2 - \sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - \bar{y})^2},$$

or

$$\begin{aligned} R^2 &= 1 - \frac{\sum_t (y_t - \hat{y}_t)^2}{\sum_t (y_t - \bar{y})^2} \\ &= 1 - \frac{\sum_t \hat{u}_t^2}{S_{yy}} = 1 - \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{S_{yy}}, \end{aligned} \tag{2.30}$$

which provides an alternative interpretation for R^2 and establishes that $0 \leq R^2 \leq 1$, so long as the underlying regression equation contains an intercept.[5] The limiting value of $R^2 = 1$ indicates perfect fit and arises if and only if $\hat{\mathbf{u}} = \mathbf{0}$ (or $\hat{u}_t = 0$, for $t = 1, 2, \ldots, T$). When T, the sample size, is finite this can only happen if the number of estimated regression coefficients, k, is equal to T. The R^2 statistic is problematic as a measure of quality of the fit of a regression models because it always increases when a new regressor is added to the model. Therefore a high value of R^2 is not by itself indicative of a good fit. An alternative measure of fit which attempts to take account of the number of estimated coefficients is due to Theil. It is called *adjusted* R^2, and written as $\bar{R}^2$:

$$\bar{R}^2 = 1 - \frac{T-1}{T-k} \frac{\sum_t \hat{u}_t^2}{\sum_t (y_t - \bar{y})^2}, \tag{2.31}$$

or equivalently (using (2.14)):

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_{YY}/(T-1)}.$$

This 'adjusted' measure provides a trade-off between fit, as measured by R^2, and parsimony as measured by $T - k$. To make this trade-off more explicit $\bar{R}^2$ is also often defined as

$$1 - \bar{R}^2 = \frac{T-1}{T-k}\left(1 - R^2\right). \tag{2.32}$$

[4] The proof is similar to that presented in Chapter 1 for the bivariate regression model and will not be repeated here.

[5] When the regression equation does not contain an intercept term, R^2 can become negative.

All the above three definitions of $\bar{R}^2$ are algebraically equivalent. Note that, unlike R^2, there is no guarantee for the $\bar{R}^2$ to be non-negative, and hence $\bar{R}$ is not always defined.

In applied econometrics, $\bar{R}^2$ is often used as a criterion of model selection. However, its use can be justified when the regression models under consideration are non-nested, in the sense that none of the models under consideration can be obtained from the others by means of some suitable parametric restrictions. In the case where the models are nested, a more suitable procedure would be to apply classical hypotheses testing procedures and test the models against one another by means of *F*- or *t*-tests. (See Chapter 3 on hypotheses testing in linear regression models.)

Remark 1 *When y_t is trended (upward or downward) it is possible to obtain an R^2 very close to unity, irrespective of whether the trend is deterministic or stochastic. This is because the denominator of R^2, namely $S_{yy} = \sum_t \left(y_t - \bar{y}\right)^2$, implicitly assumes that y_t is stationary with a constant mean and variance (see Chapter 12 for definition of stationarity). In the case of trended variables a more appropriate measure of fit would be to define R^2 with respect to the first differences of y_t, $\Delta y_t = y_t - y_{t-1}$, namely*

$$R^2_{\Delta y} = 1 - \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\sum_t \left(\Delta y_t - \overline{\Delta y}\right)^2},$$

where $\overline{\Delta y} = \sum_t \Delta y_t / T$. This measure is applicable irrespective of whether y_t is trend-stationary (namely when its deviations from a deterministic trend line are stationary), or first difference stationary. A variable is said to be first difference stationary if it must be first differenced once before it becomes stationary (see Chapter 15 for further details). The following simple relation exists between R^2 and $R^2_{\Delta y}$:

$$1 - R^2_{\Delta y} = \left(\frac{\sum_t \left(y_t - \bar{y}\right)^2}{\sum_t \left(\Delta y_t - \overline{\Delta y}\right)^2}\right)\left(1 - R^2\right).$$

Since in the case of trended y_t, for modest values of T, the sum $\sum_t \left(y_t - \bar{y}\right)^2$ will most certainly be substantially larger than $\sum_t \left(\Delta y_t - \overline{\Delta y}\right)^2$, it then follows that in practice $R^2_{\Delta y}$ will be less than R^2, often by substantial amounts. Also as T tends to infinity R^2 will tend to unity, but $R^2_{\Delta y}$ remain bounded away from unity. An alternative approach to arriving at a plausible measure of fit in the case of trended variables would be to ensure that the dependent variable of the regression is stationary by running regressions of first differences, Δy_t on the regressors, $\mathbf{x}_t$, of interest. But in that case it is important that lagged values of y_t, are also included amongst the regressors, namely a dynamic specification should be considered. This naturally leads to the analysis of error correction specifications to be discussed in Chapters 6, 23, and 24.

2.11 Partitioned regression

Consider the classical linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{2.33}$$

and suppose that $\mathbf{X}$ is partitioned into two sub-matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ of order $T \times k_1$ and $T \times k_2$ such that $k = k_1 + k_2$.[6] Partitioning $\boldsymbol{\beta}$ conformably with $\mathbf{X} = \left(\mathbf{X}_1 \vdots \mathbf{X}_2\right)$ we have

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}. \tag{2.34}$$

Such partitioned regressions arise, for example, when $\mathbf{X}_1$ is composed of seasonal dummy variables or time trends, and $\mathbf{X}_2$ contains the regressors of interest, or the 'focus' regressors. The *OLS* estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are given by the normal equations

$$\mathbf{X}_1'\mathbf{y} = \left(\mathbf{X}_1'\mathbf{X}_1\right)\hat{\boldsymbol{\beta}}_1 + \left(\mathbf{X}_1'\mathbf{X}_2\right)\hat{\boldsymbol{\beta}}_2, \tag{2.35}$$

$$\mathbf{X}_2'\mathbf{y} = \left(\mathbf{X}_2'\mathbf{X}_1\right)\hat{\boldsymbol{\beta}}_1 + \left(\mathbf{X}_2'\mathbf{X}_2\right)\hat{\boldsymbol{\beta}}_2. \tag{2.36}$$

Solving for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ we have

$$\hat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y}, \tag{2.37}$$

$$\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y}, \tag{2.38}$$

where

$$\mathbf{M}_j = \mathbf{I}_T - \mathbf{X}_j\left(\mathbf{X}_j'\mathbf{X}_j\right)^{-1}\mathbf{X}_j', \qquad \text{for} \qquad j = 1, 2.$$

The estimators of the 'focus' coefficients, $\hat{\boldsymbol{\beta}}_2$, can also be written as (recall that $\mathbf{M}_j$ are symmetric and idempotent: $\mathbf{M}_j' = \mathbf{M}_j = \mathbf{M}_j^2$):

$$\hat{\boldsymbol{\beta}}_2 = \left[\left(\mathbf{M}_1\mathbf{X}_2\right)'\left(\mathbf{M}_1\mathbf{X}_2\right)\right]^{-1}\left(\mathbf{M}_1\mathbf{X}_2\right)'\mathbf{y},$$

or

$$\hat{\boldsymbol{\beta}}_2 = \left(\widetilde{\mathbf{X}}_2'\widetilde{\mathbf{X}}_2\right)^{-1}\widetilde{\mathbf{X}}_2'\mathbf{y},$$

where $\widetilde{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{X}_2$ and $\widetilde{\mathbf{y}} = \mathbf{M}_1\mathbf{y}$ are the *residual* matrices and vectors of the regressions of $\mathbf{X}_2$ on $\mathbf{X}_1$ and of $\mathbf{y}$ on $\mathbf{X}_2$, respectively. The residuals from the regression of $\widetilde{\mathbf{y}} = \mathbf{y} - \hat{\mathbf{y}}$ on $\widetilde{\mathbf{X}}_2 = \mathbf{X}_2 - \hat{\mathbf{X}}_2$ are also given by $\widetilde{\mathbf{u}} = \widetilde{\mathbf{y}} - \widetilde{\mathbf{X}}_2\hat{\boldsymbol{\beta}}_2$. It is now easily seen that $\widetilde{\mathbf{u}}$ is in fact the same as the *OLS* residual vector from the unpartitioned regression of $\mathbf{y}$ on $\mathbf{X}$.[7] Therefore, a regression of $\mathbf{y}$ on $\widetilde{\mathbf{X}}_2$ yields the same estimate for $\boldsymbol{\beta}_2$ as the standard regression of $\mathbf{y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$ simultaneously and

[6] See Section A.9 in Appendix A for a description of partitoned matrices and their properties.

[7] Notice that

$$\begin{aligned}\tilde{\mathbf{u}} &= \left[\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right]\mathbf{y} - \left[\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right]\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 \\ &= \mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\left[\mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2\right].\end{aligned}$$

(*continued*)

without orthogonalization of the effect of $\mathbf{X}_1$ on $\mathbf{X}_2$. This property is known as the Frisch-Waugh-Lovell theorem, first introduced by Frisch and Waugh (1933), and then by Lovell (1963). For further details see also Davidson and MacKinnon (1993).

The partitioned and the unpartitioned regressions also yield the same results for the variance matrix of $\hat{\boldsymbol{\beta}}_2$. It is, therefore, possible to estimate the coefficients of the 'focus' regressors in two ways. The partitioned method first 'filters' the observations by allowing for the effect of 'non-focus' variables by running regressions of $\mathbf{y}$ on $\mathbf{X}_1$, and $\mathbf{X}_2$ on $\mathbf{X}_1$ and then computes estimates of $\boldsymbol{\beta}_2$ by regression of the filtered variables. In the case where $\mathbf{X}_1$ contains seasonal dummies, the residuals from regression of $\mathbf{y}$ on $\mathbf{X}_1$ represent seasonally adjusted $\mathbf{y}$, and similarly the residuals from regressions of the columns of $\mathbf{X}_2$ on $\mathbf{X}_1$ represent seasonally adjusted $\mathbf{X}_2$. Hence, regression of seasonally adjusted variables yields the same coefficient estimates as running a regression of seasonally unadjusted variables so long as the same seasonal dummies used to adjust $\mathbf{y}$ and $\mathbf{X}_2$ are also included in the unseasonally adjusted regression. The same results also hold for the regressions of detrended and non-detrended variables.

Special care should be exercised when using the above results from partitioned regressions. Firstly, the results do not apply when the seasonal adjustments or detrending are carried out over a time period that differs from the period over which the regression of focus variables are run. Neither do they apply if the seasonal adjustments are carried out by government agencies who often use their own in-house methods. Secondly, the computer results based on regression of seasonally adjusted variables do not generally take account of the loss in degrees of freedom associated with the estimation of seasonal or trend effects. In view of these pitfalls, it is often advisable to base estimation and hypothesis testing on the unpartitioned regression of $\mathbf{y}$ on $\mathbf{X}$. The use of partitioned regressions is helpful primarily for pedagogic purposes.

2.12 How to interpret multiple regression coefficients

The issue of how to interpret the regressions coefficients in a multiple regression model has been recently discussed in Pesaran and Smith (2014). Suppose we are interested in measuring the effects of a unit change in the regressor x_{it} on y_t. The standard procedure is to use the estimated coefficient of x_{it}, namely β_i, on the assumption that the hypothetical change in x_{it}, does not affect $x_{jt}, j \neq i$, namely it assumes that the hypothetical change in x_{it} is accompanied with holding the other regressors constant, the so called *ceteris paribus* assumption. But in almost all economic applications we are not able to control the inputs and the counterfactual exercise by which all other regressors can be held constant might not be relevant. Pesaran and Smith (2014) argue that in time series analysis, rather than focussing on the signs of individual coefficients in multiple regressions holding the other variables constant, we should measure a total impact effect which allows for direct and indirect induced changes that arise due to the historical correlations amongst the regressors. The limitation of the usual *ceteris paribus* approach lies in the fact that it ignores the stochastic interdependence of the regressors which we need to allow for in time series economic applications. Similar issues arise in the derivation of impulse response functions for

But using (2.35), $\mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_1'\mathbf{X}_1\right)\hat{\boldsymbol{\beta}}_1$, and hence

$$\begin{aligned}\tilde{\mathbf{u}} &= \mathbf{y} - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1'\mathbf{X}_1\right)\hat{\boldsymbol{\beta}}_1 \\ &= \mathbf{y} - \mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \hat{\mathbf{u}}.\end{aligned}$$

the analysis of dynamic models and have been discussed by Koop, Pesaran, and Potter (1996) and Pesaran and Shin (1998) and will be addressed in Chapter 24.

To illustrate Pesaran and Smith (2014)'s argument consider the following simple classical linear regression model with two regressors:

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t.$$

Suppose further that x_{1t} and x_{2t} are random draws from a bivariate normal distribution with the covariance matrix

$$Var\begin{pmatrix} \Delta x_{1t} \\ \Delta x_{2t} \end{pmatrix} = \begin{pmatrix} \sigma_{11} & \sigma_{21} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

It is now easily seen that

$$E(\Delta x_{2t} \,|\Delta x_{1t}) = \rho_{21} \Delta x_{1t},$$

where $\rho_{21} = \sigma_{21}/\sigma_{11}$. The total effect of a unit change in x_{it} on y_t is therefore given by

$$E(\Delta y_t \,|\Delta x_{1t}) = \left(\beta_1 + \rho_{21}\beta_1\right) \Delta x_{1t},$$

which reduces to the β_1 only if $\sigma_{21} = 0$.

As a second example, suppose that we have a quadratic function of a single regressor, so that the regression model is given by

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + u_t. \tag{2.39}$$

Here it clearly does not make any sense to ask what is the effect on y_t of a change in x_t, holding x_t^2 fixed. In this case we have

$$E(\Delta y_t \,|\Delta x_t) = \left(\beta_1 + 2\beta_2 x_t\right) \Delta x_t, \tag{2.40}$$

for sufficiently small increments, Δx_t.

Pesaran and Smith (2014) show that the total effect of a unit change in x_{it} on y_t can be estimated consistently by a simple regression of y_t on x_{it}, which is to be contrasted with the *ceteris paribus* effect of unit change in x_{it} on y_t which is given by β_i and requires estimation of the correctly specified multiple regression model.

2.13 Implications of misspecification for the *OLS* estimators

The unbiasedness property of the *OLS* estimators in the classical linear regression model crucially depends on the validity of the classical assumptions and the correct specification of the regression equation. Here we consider the effects of misspecification, that results from adding or omitting a regressor in error, on the *OLS* estimators. In Chapter 3 we consider the implications of such misspecifications for inference on the regression coefficients.

2.13.1 The omitted variable problem

Suppose that y_t's are generated according to the classical linear regression equation

$$y_t = \alpha + \beta_1 x_t + \beta_2 z_t + u_t, \tag{2.41}$$

but the investigator estimates the simple regression equation

$$y_t = \alpha + \beta x_t + \varepsilon_t, \tag{2.42}$$

which omits the regressor z_t. The new error, ε_t, now contains the effect of the omitted variable and the orthogonality assumption that requires x_t and ε_t to be uncorrelated might no longer hold. To see this consider the *OLS* estimator of β in (2.42), which is given by

$$\hat{\beta} = \frac{\sum_{t=1}^{T} (x_t - \bar{x}) \left(y_t - \bar{y}\right)}{\sum_{t=1}^{T} (x_t - \bar{x})^2}.$$

Under the correct model (33.34)

$$y_t - \bar{y} = \beta_1 (x_t - \bar{x}) + \beta_2 (z_t - \bar{z}) + u_t - \bar{u}.$$

Hence

$$\hat{\beta} = \beta_1 + \beta_2 \frac{\sum_t (x_t - \bar{x}) (z_t - \bar{z})}{\sum_t (x_t - \bar{x})^2} + \frac{\sum_t (x_t - \bar{x}) (u_t - \bar{u})}{\sum_t (x_t - \bar{x})^2},$$

and taking expectations conditional on the regressors

$$E\left(\hat{\beta}\right) = \beta_1 + \beta_2 b_{x \bullet z}, \tag{2.43}$$

where $b_{x \bullet z}$ stands for the *OLS* estimator of the regression coefficient of x_t on z_t. In general, therefore, $\hat{\beta}$ is not an unbiased estimator of β_1 [the 'true' regression coefficient of x_t in (2.41)]. The extent of the bias depends on the importance of the z_t variable as measured by β_2 and the degree of the dependence of x_t on z_t. Only in the case where x_t and z_t are uncorrelated $\hat{\beta}$ will yield an unbiased estimator of β_1. See Section 3.13 on the effects of omitting relevant regressors on testing hypothesis involving the regression coefficients.

The omitted regressor bias can be readily generalized to the case where two or more relevant regressors are omitted. The appropriate set up is the partitioned regression equation given in (2.34). Suppose that in that equation the regressors $\mathbf{X}_2$ are incorrectly omitted and $\boldsymbol{\beta}_1$ is estimated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_1' \mathbf{X}_1)^{-1} \mathbf{X}_1' \mathbf{y}.$$

Then, under (2.34), it is easily seen that

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_1 \,|\mathbf{X}) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{P}_{12}\boldsymbol{\beta}_2,$$

Also see Exercise 3 at the end of this chapter.

2.13.2 The inclusion of irrelevant regressors

Inclusion of irrelevant regressors in the regression equation is less problematic. For example, suppose that the correct model is

$$y_t = \alpha + \beta x_t + u_t,$$

but we estimate the expanded regression equation by mistake:

$$y_t = \alpha + \beta_1 x_t + \beta_2 z_t + \varepsilon_t.$$

The *OLS* estimator of β_1 in this regression will still be unbiased, but will no longer be an efficient estimator. There will also be the possibility of a multicollinearity problem that can arise if the erroneously included regressor, z_t, is highly correlated with x_t (see Section 15.3.1). In general suppose that the correct regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{2.44}$$

but $\boldsymbol{\beta}$ is estimated by running the expanded regression of $\mathbf{y}$ on $\mathbf{X}$ and $\mathbf{Z}$. The *OLS* estimator of the coefficients of $\mathbf{X}$ in this regression, say $\boldsymbol{\beta}_1$, is given by (see also (2.37))

$$\hat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}'\mathbf{M}_z\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{M}_z\mathbf{y},$$

where $\mathbf{M}_z = \mathbf{I}_T - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}$. Under (2.44) we have

$$E\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta} \,|\mathbf{X}, \mathbf{Z}\right) = \left(\mathbf{X}'\mathbf{M}_z\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{M}_z E\left(\mathbf{u} \,|\mathbf{X}, \mathbf{Z}\right).$$

Therefore, so long as $\mathbf{Z}$ as well as $\mathbf{X}$ are strictly exogenous and the orthogonality assumption $E\left(\mathbf{u} \,|\mathbf{X}, \mathbf{Z}\right) = \mathbf{0}$ is satisfied we obtain

$$E\left(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta} \,|\mathbf{X}, \mathbf{Z}\right) = \mathbf{0},$$

or unconditionally

$$E\left(\hat{\boldsymbol{\beta}}_1\right) = \boldsymbol{\beta}.$$

Notice, however, that the additional variables in $\mathbf{Z}$ can not be weakly exogenous. For example, adding lagged values of y_t to the regressors in error can lead to biased estimators.

2.14 Linear regressions that are nonlinear in variables

A linear regression model does not necessarily require the relationship between **y** (the regressand) and **x** (the regressor) to be linear. A simple example of a linear regression model with a nonlinear functional relationship between **y** and **x** is given by the quadratic regression equation:

$$y_i = \alpha + \beta x_i + \gamma x_i^2 + u_i.$$

To transform this nonlinear relation to a linear regression model, set $z_i = x_i^2$ and write the quadratic equation as

$$y_i = \alpha + \beta x_i + \gamma z_i + u_i,$$

which is a linear regression in the two regressors x_i and z_i. Other examples of nonlinear relations that are transformable to linear regressions are general polynomial regressions, logistic models, log-linear, semi-log-linear and inverse models. Here we examine some of these models in more detail.

Example 2 *Consider the following Cobb–Douglas production function*

$$Q_i = AL_i^{\alpha} K_i^{\beta} \exp(u_i),$$

where Q_i is output of firm i, L_i and K_i are the quantities of labour and capital used in the production process, and u_i are independently distributed productivity shocks. Taking logarithms of both sides now yields the linear logarithmic specification

$$\log Q_i = \log A + \alpha \log L_i + \beta \log K_i + u_i,$$

and setting $y_i = \log Q_i$, $x_{1i} = \log L_i$, $x_{2i} = \log K_i$, $a = \log A$, yields

$$y_i = a + \alpha x_{1i} + \beta x_{2i} + u_i,$$

which is a linear regression equation in the two regressors x_{1i} and x_{2i}. The estimate of A can now be obtained by $\hat{A} = \exp(\hat{a})$, where $\hat{a}$ is the OLS estimate of the intercept term in the above regression.

Example 3 (Logistic function with a known saturation level) *The logistic model has the general form*

$$Y_i = \frac{A}{1 + \gamma x_i^{\beta} \exp(u_i)}, \qquad \beta, \gamma > 0, \qquad x_i > 0,$$

where A is the saturation level of Y, which is assumed to be known. We also assume that $A > Y_i$, for all i. This is clearly a nonlinear model in terms of Y and x. To transform this model into a linear regression model in terms of the unknown parameters γ and β, we first note that

$$\gamma x_i^{\beta} \exp(u_i) = \frac{A}{Y_i} - 1,$$

which upon taking logarithms of both sides yields

$$y_i = \log\left(\frac{A - Y_i}{Y_i}\right) = \alpha + \beta \log x_i + u_i,$$

in which $\alpha = \log(\gamma)$. *In the case where A is known the parameters* α *and* β *(and hence* γ *and* β*) can be estimated by the OLS regression of* $y_i = \log\left(\frac{A-Y_i}{Y_i}\right)$ *on log*x_i. *The logistic function, has important applications in econometrics (e.g. ownership of demand for durable goods, TV's, cars etc.) and in population studies.*

Other examples of nonlinear functions that can be transformed into linear regressions include semi-logarithmic model

$$y_i = \alpha + \beta \log x_i + u_i,$$

and the inverse model

$$y_i = \alpha + \frac{\beta}{x_i} + u_i.$$

These models have proved very useful in cross-section studies of household consumption behaviour.

2.15 Further reading

Further reading on multiple regression and on the properties of *OLS* estimator can be found in Wooldridge (2000) and in Greene (2002) (see Chapters 2–4). An interesting geometric interpretation of linear regression, shedding light on the numerical properties of *OLS*, is presented in Davidson and MacKinnon (1993). The latter also provides an in-depth discussion on the Frisch–Waugh–Lovell theorem, and partitioned regression.

2.16 Exercises

1. Suppose that in the classical regression model $y_i = \alpha + \beta x_i + u_i$ the true value of the constant, α, is zero. Compare the variance of the *OLS* estimator for β computed without a constant term with that of the *OLS* estimator for β computed with the constant term.
2. Consider the following linear regression model

$$y_t = \alpha + \beta x_t + \gamma w_t + u_t. \tag{2.45}$$

Suppose that the classical assumptions are applicable to (2.45), but β is estimated by running an *OLS* regression of y_t on a vector of ones and x_t. Denote such an estimator by $\widehat{\beta}$, and show that $\widehat{\beta}$ is a biased estimator of β in (2.45). Derive the formula for the bias of $\widehat{\beta}$ in terms of the correlation coefficient of x_t and w_t, and their variances, namely $\rho_{xw}, \sigma_x^2, \sigma_w^2$.

3. Consider the following partitioned classical linear regression model:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u},$$

where y is a $T \times 1$ vector of observations on the dependent variable, and $\mathbf{X}_1$ and $\mathbf{X}_2$ are $T \times k_1$ and $T \times k_2$ observation matrices on the regressors.

(a) Show that if we omit the variables included in $\mathbf{X}_2$, and estimate $\boldsymbol{\beta}_1$ by running a regression of $\mathbf{y}$ on $\mathbf{X}_1$ only, then $\hat{\boldsymbol{\beta}}_1$ is generally biased with the bias:

$$E(\hat{\boldsymbol{\beta}}_1|\mathbf{X}) - \boldsymbol{\beta}_1 = \mathbf{P}_{12}\boldsymbol{\beta}_2, \quad \text{where } \mathbf{P}_{12} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2,$$

where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$.

(b) Interpret the elements of matrix $\mathbf{P}_{12}$. Under what conditions $\hat{\boldsymbol{\beta}}_1$ will be unbiased?

(c) A researcher is estimating the demand equation for furniture using cross-section data. As regressors she uses an intercept term, the relative price of furniture, and omits the relevant income variable. Find an expression for the bias of the *OLS* estimate of the price variable in such a regression. What other regressors should she have considered, and how could their omission have affected her estimate of the price effect?

4. Consider the following linear regression model

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \tag{2.46}$$

and suppose that the observations (y_t, x_{1t}, x_{2t}), for $t = 1, 2, \ldots, T$ are available.

(a) Specify the assumptions under which (2.46) can be viewed as a classical linear regression model. In your response clearly distinguish between the cases where x_{1t} and x_{2t} are fixed in repeated samples, strictly exogenous, and weakly exogenous (see Chapter 9 for definition of strictly exogenous, and weakly exogenous regressors).

(b) Suppose that the classical assumptions are applicable to (2.46), but β_1 is estimated by running an *OLS* regression of y_t on a vector of ones and x_{1t}, and β_2 is estimated by running an *OLS* regression of y_t on a vector of ones and x_{2t}. Denote these estimators by $\widehat{\beta}_{yx_1}$ and $\widehat{\beta}_{yx_2}$. Show that in general $\widehat{\beta}_{yx_1}$ and $\widehat{\beta}_{yx_1}$ are biased estimators of β_1 and β_2 in (2.46).

(c) Denote the *OLS* estimators of β_1 and β_2 in the regression of y_t on x_{1t} and x_{2t} as in (2.46) by $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. Show that

$$\hat{\beta}_1 = \frac{\widehat{\beta}_{yx_1} - r(s_2/s_1)\widehat{\beta}_{yx_2}}{1 - r^2},$$
$$\hat{\beta}_2 = \frac{\widehat{\beta}_{yx_2} - r(s_1/s_2)\widehat{\beta}_{yx_1}}{1 - r^2},$$

where s_1 and s_2 are the standard deviations of x_{1t} and x_{2t}, respectively, and r denotes the correlation coefficients of x_{1t} and x_{2t}. Discuss the relevance of these results for empirical time series research.

5. Consider the regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \qquad \mathbf{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_T),$$

where $\mathbf{X}$ is a $T \times k$ stochastic matrix of rank k, distributed independently of $\mathbf{u} = (u_1, u_2, \ldots, u_T)'$, and $u_t \backsim IID(0, \sigma^2)$.

(a) Let $\lambda_{\max}(\mathbf{X}'\mathbf{X})$ and $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ denote the largest and the smallest characteristic roots (or eigenvalues) of $\mathbf{X}'\mathbf{X}$. Prove that the following four statements are equivalent:
- $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ tends to infinity
- $\lambda_{\max}\left((\mathbf{X}'\mathbf{X})^{-1}\right)$ tends to zero
- *Trace* $\left((\mathbf{X}'\mathbf{X})^{-1}\right)$ tends to zero
- Every diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ tends to zero

(b) Using the results under (*a*), or otherwise show that the *OLS* estimator of $\boldsymbol{\beta}$ is consistent if $\lambda_{\min}(\mathbf{X}'\mathbf{X})$ tends to infinity.

(c) Prove $\hat{\sigma}^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/T$ is a consistent estimator of σ^2, where $\hat{\mathbf{u}}$ is the vector of *OLS* residuals.

3 Hypothesis Testing in Regression Models

3.1 Introduction

Statistical hypothesis testing is at the core of the classical theory of statistical inference. Although it is closely related to the problem of estimation, it can be considered almost independently of it. In this chapter, we introduce some key concepts of statistical inference, and show their use to investigate the statistical significance of the (linear) relationships modelled through regression analysis, or to investigate the validity of the classical assumptions in simple and multiple linear regression.

3.2 Statistical hypothesis and statistical testing

A *statistical hypothesis* is an assertion about the distribution of one or more random variables. If the hypothesis completely specifies the probability distribution, it is called a *simple* hypothesis, otherwise it is called a *composite* hypothesis. For example, suppose $x_1, x_2, \ldots, x_T$ are drawn from $N(\theta, 1)$. Then $H : \theta = 0$ is a simple hypothesis, while $H : \theta > 0$ is a composite hypothesis. If one hypothesis can be derived as a limiting sequence of another, we say the two hypotheses are *nested*. If neither hypothesis can be obtained from the other as a limiting process, then we call the hypotheses under consideration *non-nested*. For example, suppose $x_1, x_2, \ldots, x_T$ are drawn from log-normal distribution under H_0, while under H_1 they are drawn from an exponential distribution. Then H_0 and H_1 are non-nested hypotheses. We refer to Chapter 11 for a review of tests for non-nested hypotheses.

3.2.1 Hypothesis testing

A test of a statistical hypothesis H is a rule for *rejecting* H. If the sample space is denoted by $\chi = (x_1, x_2, \ldots, x_T)$, a test procedure decomposes χ into two regions. If $(x_1, x_2, \ldots, x_T) \in C_T$, where C_T is called the critical or *rejection region* of the test, then H is rejected, otherwise H is not rejected. In practice we often map $(x_1, x_2, \ldots, x_T)$ into a test statistic $T(x_1, x_2, \ldots, x_T)$ and consider whether $T(x_1, x_2, \ldots, x_T) \geq C_T$ or not.

The hypothesis being tested (i.e. the maintained hypothesis) is usually denoted by H_0 and is called the *null* hypothesis. The hypothesis against which H_0 is tested is called the *alternative* hypothesis and is usually denoted by H_1.

3.2.2 Types of error and the size of the test

The decision rule yields two types of error:

- The *type I error* is the error involved in rejecting H_0 when it is true
- The *type II error* is the error involved in not rejecting H_0 when it is false

The probability of a type I error is called the *size of the test* and, often denoted by α_T, $\alpha_T \times 100$ per cent, is also called the *significance level* of the test. The probability of the type II error is called the *size of the type II error* and is often denoted by β_T. Ideally, we would like both errors to be as small as possible. However, there is a trade-off between the two, and by reducing the probability of a type I error, we must increase the probability of a type II error.

The *power* of a test is defined as 1 minus the size of the type II error, namely $power_T = 1 - \beta_T$. For a given significance level, α_T, we would like the power of the test, $power_T$, to be as large as possible.

Example 4 (Testing a hypothesis about a mean) *Assume we have a sample of T observations $x_1, x_2, \ldots, x_T$, obtained as random draws from a normal $N(\mu, \sigma^2)$ distribution, with σ^2 known. Suppose that we wish to test $H_0 : \mu = \mu_0$, where μ_0 is a given (assumed) value of μ. To this end, consider the sample mean $\bar{x} = T^{-1} \sum_{i=1}^{T} x_i$. Under the null hypothesis the random variable*

$$z = \frac{\sqrt{T}\bar{x} - \mu_0}{\sigma},$$

is distributed as a $N(0, 1)$ and the critical values of the normal distribution will be applicable. Setting the significance level at 5 per cent, the critical value for a two-sided test (with the alternative being $\mu \neq \mu_0$) is 1.96. Hence, in this case the power of the test is the probability that the absolute value of the test statistic will exceed 1.96 given that the true value of μ is not μ_0. The power clearly depends on the alternative value selected for μ. As expected, the test becomes more powerful the further the true mean is from the hypothesized value. The interval

$$P\left(\bar{x} - 1.96\sigma/\sqrt{T} \leq \mu \leq \bar{x} + 1.96\sigma/\sqrt{T}\right) = 0.95,$$

is called the 95 per cent confidence interval of μ.

Let the critical region of a test be defined by $T(x_1, x_2, \ldots, x_T) \geq C_T$, we have

$$\begin{aligned} \text{Prob. of type I error} &= \Pr\{T(x_1, x_2, \ldots, x_T) \geq C_T \,|H_0\} = \alpha_T, \\ \text{Prob. of type II error} &= \Pr\{T(x_1, x_2, \ldots, x_T) < C_T \,|H_1\} = \beta_T. \end{aligned}$$

Let Π_T denote the power of the test, then

$$\Pi_T = 1 - \beta_T = 1 - \Pr\{T(x_1, x_2, \ldots, x_T) < C_T | H_1\},$$

or equivalently,

$$\Pi_T = \Pr\{T(x_1, x_2, \ldots, x_T) \geq C_T | H_1\}.$$

3.3 Hypothesis testing in simple regression models

In deriving the ordinary least squares (OLS) estimator and its properties in Chapter 2, we have not used Assumption A5 on the normality of u_t. This assumption is useful for hypotheses testing. Consider first the simple regression model

$$y_t = \alpha + \beta x_t + u_t,$$

and assume that Assumptions A1–A4 hold (see Chapter 2), together with Assumption A5, that is, $u_t \sim N(0, \sigma^2)$. Suppose that we are interested in testing the null hypothesis

$$H_0 : \beta = \beta_0,$$

against the two-sided alternative hypothesis

$$H_1 : \beta \neq \beta_0,$$

where β_0 is a given value of β. To construct a test for β, first recall that, from (1.25) and (1.26),

$$\hat{\beta} = \sum_{t=1}^{T} w_t y_t,$$

where

$$w_t = \frac{x_t - \bar{x}}{\sum_{s=1}^{T} (x_s - \bar{x})^2}.$$

Replacing $y_t = \alpha + \beta x_t + u_t$ in the above expression now yields

$$\hat{\beta} = \sum_{t=1}^{T} w_t (\alpha + \beta x_t + u_t),$$

$$\hat{\beta} = \alpha \left(\sum_{t=1}^{T} w_t \right) + \beta \left(\sum_{t=1}^{T} w_t x_t \right) + \sum_{t=1}^{T} w_t u_t,$$

and since $\sum_{t=1}^{T} w_t = 0$, $\sum_{t=1}^{T} w_t x_t = 1$ (see the derivations in Section 1.9), we have

$$\hat{\beta} = \beta + \sum_{t=1}^{T} w_t u_t. \tag{3.1}$$

Noting that the weighted average of normal variates is also normal, it follows that

$$\hat{\beta} \,|\mathbf{x} \sim N\left[\beta, Var\left(\hat{\beta}\right)\right], \tag{3.2}$$

where

$$Var\left(\hat{\beta}\right) = \sigma^2 \sum_{t=1}^{T} w_t^2 = \frac{\sigma^2}{\sum_{t=1}^{T} (x_t - \bar{x})^2}.$$

In the case where σ^2 is known, we can base the test of $H_0 : \beta = 0$, on the following standardized statistic

$$Z_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{\sqrt{Var\left(\hat{\beta}\right)}} = \frac{\hat{\beta} - \beta_0}{S.E.\left(\hat{\beta}\right)}, \tag{3.3}$$

where *S.E.* $(\cdot)$ stands for the standard errors. Under the null hypothesis, $Z_{\hat{\beta}} \sim N(0, 1)$ and the critical values of the normal distribution will be applicable.

The appropriate choice of the critical values depends on the distribution of the test statistic, the size of the test (or the level of significance), and whether the alternative hypothesis is two sided, (namely $H_1 : \beta \neq \beta_0$) or one-side, namely whether $H_1 : \beta \geq \beta_0$ or $H_1 : \beta \leq \beta_0$.

In the case where σ^2 is not known, the use of statistic $Z_{\hat{\beta}}$ defined by (3.3) is not feasible and σ^2 needs to be replaced by its estimate. Using the unbiased estimator of σ^2, given by (1.34), namely

$$\hat{\sigma}^2 = \frac{\sum_t \left(y_t - \hat{\alpha} - \hat{\beta} x_t\right)^2}{T - 2},$$

we have the *t*-statistic

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{\sqrt{\widehat{Var}\left(\hat{\beta}\right)}} = \frac{\hat{\beta} - \beta_0}{\hat{\sigma} \Big/ \left\{\sum_t (x_t - \bar{x})^2\right\}^{\frac{1}{2}}},$$

which under the null hypothesis, $H_0 : \beta = \beta_0$ has a *t*-distribution with $T - 2$ degrees of freedom. The $t_{\hat{\beta}}$ statistic is pivotal in the sense that it does not depend on any unknown parameters.

Example 5 *Suppose we are interested to test the hypothesis that the marginal propensity to consume out of disposable income is equal to unity. Using aggregate UK consumption data over the period 1948–89 we obtained the following OLS estimates:*

$$\begin{array}{rccc} \hat{c}_t = & 7600.3 & + & 0.87233 \quad y_t. \\ & (2108.9) & & (0.01169) \end{array}$$

The bracketed figures are standard errors. The estimate of the marginal propensity to consume is equal to $\hat{\beta} = 0.87233$. *To test* $H_0 : \beta = 1$ *against* $H_1 : \beta \neq 1$ *we compute the t-statistic*

$$t_{\hat{\beta}} = \frac{\hat{\beta} - \beta_0}{S.E.(\hat{\beta})} = \frac{0.87233 - 1.0}{0.01169} = -10.92.$$

The number of degrees of freedom of this test is equal to $42 - 2 = 40$, *and the* 95 *per cent critical value of the t-distribution with* 40 *degrees of freedom for a two-sided test is equal to* ± 2.021. *Hence since the value of* $t_{\hat{\beta}}$ *for the test of* $\beta = 1$ *against* $\beta \neq 1$ *is well below the critical value of the test (i.e.,* -2.021*) we reject the null hypothesis that* $\beta = 1$.

3.4 Relationship between testing $\beta = 0$, and testing the significance of dependence between *Y* and *X*

Recall that the correlation coefficient between Y and X is estimated by (see Section 1.9)

$$\hat{\rho}^2_{XY} = \frac{S^2_{XY}}{S_{XX} S_{YY}}.$$

But since

$$\hat{\beta} = \frac{S_{XY}}{S_{XX}},$$
$$\widehat{Var}(\hat{\beta}) = \frac{\hat{\sigma}^2}{S_{XX}},$$

we have

$$\hat{\rho}^2_{XY} = \frac{\hat{\beta}^2 S^2_{XX}}{S_{XX} S_{YY}} = \frac{\hat{\beta}^2 S_{XX}}{S_{YY}}. \tag{3.4}$$

The t-statistic for testing $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ is given by

$$\hat{t}_\beta = \frac{\hat{\beta}}{\sqrt{\widehat{Var}(\hat{\beta})}},$$

or upon using the above results:

$$\hat{t}_{\hat{\beta}}^2 = \frac{\hat{\beta}^2 S_{XX}}{\hat{\sigma}^2}. \tag{3.5}$$

Finally, recall from the decomposition of $S_{YY} = \sum \left(y_t - \bar{y}\right)^2$ in the analysis of variance table that (see Section 1.5)

$$\hat{\rho}_{XY}^2 = 1 - \frac{\sum_t \left(y_t - \hat{y}_t\right)^2}{\sum_t \left(y_t - \bar{y}\right)^2} = 1 - \frac{(T-2)\,\hat{\sigma}^2}{S_{YY}},$$

or

$$\hat{\sigma}^2 = \frac{S_{YY}\left(1 - \hat{\rho}_{XY}^2\right)}{T-2}. \tag{3.6}$$

Consequently, using (3.4) and (3.5) in (3.6) we have

$$t_{\hat{\beta}}^2 = \frac{(T-2)\,\hat{\rho}_{XY}^2}{\left(1 - \hat{\rho}_{XY}^2\right)}. \tag{3.7}$$

Alternatively, $\hat{\rho}_{XY}^2$ can be written as an increasing function of $t_{\hat{\beta}}^2$ for $T > 2$, namely

$$\hat{\rho}_{XY}^2 = \frac{t_{\hat{\beta}}^2}{T - 2 + t_{\hat{\beta}}^2} < 1. \tag{3.8}$$

These results show that in the context of a simple regression model the statistical test of the 'fit' of the model (i.e., $H_0 : \rho_{XY} = 0$ against $H_1 : \rho_{XY} \neq 0$) is the *same* as the test of zero restriction on the slope coefficient of the regression model (i.e., test of $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$). Moreover, the test results under the null hypothesis of a zero relationship between Y and X is equivalent to testing the significance of the reverse regression of X on Y, namely testing $H_0 : \delta = 0$, against $H_1 : \delta \neq 0$, in the reverse regression

$$x_t = a_x + \delta y_t + v_t, \tag{3.9}$$

assuming that the classical assumptions now apply to this model. Of course, it is clear that the classical assumptions cannot apply to the regression of Y on X and to the reverse regression of X on Y at the same time. But testing the null hypothesis that $\beta = 0$ and $\delta = 0$ are equivalent since the null states that there is no relationship between the two variables. However, if the null of no relationship between Y and X is rejected, then to measure the size of the effect of X on Y ($\beta_{X \cdot Y}$) as compared with the size of the effect of Y on X ($\beta_{Y \cdot X}$), will crucially depend on whether the classical assumptions are likely to hold for the regression of Y on X or for the reverse regression of X on Y. As was already established in Chapter 1, $\hat{\beta}_{Y \cdot X}\hat{\beta}_{X \cdot Y} = \hat{\rho}_{YX}^2 = \hat{\rho}_{XY}^2$ (see (1.9)), from

which it follows in general that the estimates of the effects of X on Y and the effects of Y on X do not match, in the sense that $\hat{\beta}_{Y\cdot X}$ is not equal to $1/\hat{\beta}_{X\cdot Y}$, unless $\hat{\rho}^2_{XY} = 1$, which does not apply in practice.

Hence, in order to find the size of the effects the direction of the analysis (whether Y is regressed on X or X regressed on Y) matters crucially. But, if the purpose of the analysis is simply to test for the significance of the statistical relationship between Y and X, the direction of the regression does not matter and it is sufficient to test the null hypothesis of zero correlation (or more generally zero dependence) between Y and X. This can be done using a number of alternative measures of dependence between Y and X. In addition to ρ_{YX}, one can also use Spearman rank correlation and Kendall's τ coefficients defined in Section 1.4. The rank correlation measures are less sensitive to outliers and are more appropriate when the underlying bivariate distribution of (Y and X) show significant departures from Gaussianity and the sample size, T, under consideration is small. But in cases where T is sufficient large (60 or more), and the underlying bivariate distribution has fourth-order moments, then the use of simple correlation coefficient, ρ_{YX}, seems appropriate and tests based on it are likely to be more powerful than tests based on rank correlation coefficients.

Under the null hypothesis that Y and X are independently distributed $\sqrt{T}\hat{\rho}_{YX}$ is asymptotically distributed as $N(0, 1)$, and a test of $\rho_{YX} = 0$ can be based on

$$z_\rho = \sqrt{T}\hat{\rho}_{YX} \rightarrow_d N(0, 1).$$

Fisher has derived an exact sample distribution for $\hat{\rho}_{YX}$ when the observations are from an underlying bivariate normal distribution. But in general no exact sampling distribution is known for $\hat{\rho}_{YX}$ in the case of non-Gaussian processes. In small samples more accurate inferences can be achieved by basing the test of $\rho_{YX} = 0$ on $t_{\hat{\beta}} = \hat{\rho}_{YX}\sqrt{(T-2)/(1-\hat{\rho}^2_{YX})}$ which is distributed approximately as the Student's t with $T - 2$ degrees of freedom. This result follows from the equivalence of testing $\rho_{YX} = 0$ and testing $\beta = 0$ in the simple regression model $y_t = \alpha + \beta x_t + u_t$.

To use the Spearman rank correlation to test the null hypothesis that Y and X are independent, we recall from (1.10) that the Spearmen rank correlation, r_s, between Y and X is defined by

$$r_s = 1 - \frac{6\sum_{t=1}^{T} d_t^2}{T(T^2-1)}, \tag{3.10}$$

where d_t is the difference between the ranks of the two variables. Under the null hypothesis of zero rank correlation between y and x ($\rho_s = 0$, where ρ_s is the rank correlation coefficient in the population from which the sample is drawn) we have

$$Var(r_s) = \frac{1}{T-1}. \tag{3.11}$$

Furthermore, for sufficiently large T, r_s is normally distributed. A more accurate approximate test of $\rho_s = 0$ is given by

$$t_{s,T-2} = \frac{r_s\sqrt{T-2}}{\sqrt{1-r_s^2}}, \tag{3.12}$$

which is distributed (under $\rho_s = 0$) as Student t with $T - 2$ degrees of freedom

Alternatively, Kendall's τ correlation coefficient, defined by (1.11), can be used to test the null hypothesis that Y and X are independent, or in the context of Kendall's measure under the null hypothesis of zero concordance between Y and X in the population. Under the null of zero concordance $E(\tau_T) = 0$ and $Var(\tau_T) = 2(2T+5)/[9T(T-1)]$, and the test can be based on

$$z_\tau = \frac{\sqrt{9T(T-1)}\tau_T}{\sqrt{2(2T+5)}}, \tag{3.13}$$

which is approximately distributed as $N(0, 1)$.

3.5 Hypothesis testing in multiple regression models

Consider now the multiple regression model

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + u_t, \qquad u_t \sim N\left(0, \sigma^2\right), \tag{3.14}$$

and suppose that we are interested in testing the null hypothesis on the j^{th} coefficient

$$H_0 : \beta_j = \beta_{j0}, \tag{3.15}$$

against the two-sided alternative

$$H_1 : \beta_j \neq \beta_{j0}.$$

Using a similar line of reasoning as above, it is easy to see that conditional on $\mathbf{X}$

$$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj}\right),$$

where $\left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj}$ is the (j, j) element of the matrix $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ (see expression (2.13)). Hence, in the case where σ^2 is known, the test can be based on the following standardized statistic

$$Z_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_{j0}}{\sigma\left[(\mathbf{X}'\mathbf{X})^{-1}_{jj}\right]^{1/2}},$$

Under the null hypothesis (3.15), $Z_{\hat{\beta}_j} \sim N(0, 1)$ and the critical values of the normal distribution will be applicable. When σ^2 is not known, the unbiased estimator of σ^2, given by (2.14), namely

$$\hat{\sigma}^2 = \frac{\sum_t \hat{u}_t^2}{T-k} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{T-k},$$

can be used, where k is the number of regression coefficients (inclusive of an intercept, if any). Replacing σ^2 with $\hat{\sigma}^2$, yields the t-statistic

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{\sigma}\left[(\mathbf{X}'\mathbf{X})^{-1}_{jj}\right]^{1/2}},$$

which, under the null hypothesis, H_0 has a t-distribution with $T - k$ degrees of freedom.

3.5.1 Confidence intervals

Knowledge of the distribution of the estimated regression coefficients $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ can also be used to construct exact confidence intervals for the regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$. Consider the multiple regression model (3.14), and suppose that we are interested in $(1 - \alpha) \times 100$ per cent confidence interval for the regression coefficients. Since $\hat{\beta}_j$ individually have a t-distribution with $T - k$ degree of freedom, then the $(1 - \alpha) \times 100$ per cent confidence interval for β_j is given by

$$\hat{\beta}_j \pm t_\alpha\,(T - k)\,\widehat{S.E.}\left(\hat{\beta}_j\right), \tag{3.16}$$

where $t_\alpha\,(T - k)$ is the $(1 - \alpha) \times 100$ per cent critical value of the t-distribution with $T - k$ degrees of freedom for a two-sided test, and $\widehat{S.E.}\left(\hat{\beta}_j\right)$ is the estimated standard error of $\hat{\beta}_j$.

3.6 Testing linear restrictions on regression coefficients

Consider the linear regression model

$$y_t = \alpha + \beta_1 x_{t1} + \beta_2 x_{t2} + u_t, \tag{3.17}$$

and assume that it satisfies all the classical assumptions. Suppose now that we are interested in testing the hypothesis

$$H_0 : \beta_1 + \beta_2 = 1,$$

against

$$H_1 : \beta_1 + \beta_2 \neq 1.$$

Let

$$\delta = \beta_1 + \beta_2 - 1, \tag{3.18}$$

then the test of H_0 against H_1 simplifies to the test of

$$H_0 : \delta = 0,$$

against

$$H_1 : \delta \neq 0.$$

The *OLS* estimator of δ is given by

$$\hat{\delta} = \hat{\beta}_1 + \hat{\beta}_2 - 1,$$

and the relevant statistic for testing $\delta = 0$ is given by

$$t_{\hat{\delta}} = \frac{\hat{\delta} - 0}{\sqrt{\widehat{Var}\left(\hat{\delta}\right)}} = \frac{\hat{b}_1 + \hat{b}_2 - 1}{\sqrt{\widehat{Var}\left(\hat{\delta}\right)}}.$$

where

$$\widehat{Var}\left(\hat{\delta}\right) = \widehat{Var}\left(\hat{\beta}_1\right) + \widehat{Var}\left(\hat{\beta}_2\right) + 2\widehat{\text{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right).$$

The relevant expressions of the variance-covariance matrix of the regression coefficients are given in relations (2.20)–(2.22).

An alternative procedure for testing $\delta = 0$ which does not require knowledge of Cov $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ would be to use (3.18) to solve for β_1 or β_2 in the regression equation (3.17). Solving for β_2, for example, we have

$$y_t = \beta_0 + \beta_1 x_{t1} + \left(\delta - \beta_1 + 1\right) x_{t2} + u_t,$$

or

$$y_t - x_{t2} = \beta_0 + \beta_1 \left(x_{t1} - x_{t2}\right) + \delta x_{t2} + u_t. \tag{3.19}$$

Therefore, the test of $\delta = 0$ against $\delta \neq 0$ can be carried out by means of a simple t-test on the regression coefficient of x_{t2} in the regression of $(y_t - x_{t2})$ on $(x_{t1} - x_{t2})$ and x_{t2}.

Example 6 *This example describes two different methods of testing the hypothesis of constant returns to scale in the context of the Cobb–Douglas (CD) production function*

$$Y_t = AK_t^{\alpha} L_t^{\beta} e^{u_t}, \quad t = 1, 2, \ldots, T, \tag{3.20}$$

where Y_t = Output, K_t = Capital Stock, L_t = Employment. The unknown parameters A, α and β are fixed, and u_ts are serially uncorrelated disturbances with zero means and a constant variance. We also assume that u_ts are distributed independently of K_t and L_t. The constant returns to scale hypothesis postulates that proportionate changes in inputs (K_t and L_t) result in the same proportionate change in output. For example, doubling K_t and L_t should, under the constant returns to scale hypothesis, lead also to the doubling of Y_t. This imposes the following parametric restriction on (3.20):

$$H_0: \quad \alpha + \beta = 1,$$

which we consider as the null hypothesis and derive an appropriate test of it against the two-sided alternative:

$$H_1: \quad \alpha + \beta \neq 1.$$

In order to implement the test of H_0 against H_1, we first take logarithms of both sides of (3.20), which yield the log-linear specification

$$LY_t = a + \alpha LK_t + \beta LL_t + u_t \tag{3.21}$$

where

$$LY_t = \log(Y_t), \quad LK_t = \log(K_t), \quad LL_t = \log(L_t)$$

and $a = \log(A)$. It is now possible to obtain estimates of α and β by running OLS regressions of LY_t on LK_t and LL_t (for $t = 1, 2, \ldots, T$), including an intercept in the regression. Denote the OLS estimates of α and β by $\hat{\alpha}$ and $\widehat{\beta}$, and define a new parameter, δ, as

$$\delta = \alpha + \beta - 1. \tag{3.22}$$

The hypothesis $\alpha + \beta = 1$ against $\alpha + \beta \neq 1$ can now be written equivalently as

$$\begin{aligned} H_0: &\quad \delta = 0, \\ H_1: &\quad \delta \neq 0. \end{aligned}$$

We now consider two alternative methods of testing $\delta = 0$: a direct method and a regression method. The first method directly focuses on the OLS estimates of δ, namely $\hat{\delta} = \hat{\alpha} + \hat{\beta} - 1$, and examines whether this estimate is significantly different from zero. For this we need an estimate of the variance of $\hat{\delta}$. We have

$$V(\hat{\delta}) = V(\hat{\alpha}) + V(\hat{\beta}) + 2\,\mathrm{Cov}\left(\hat{\alpha}, \hat{\beta}\right),$$

where $V(\cdot)$ and $\mathrm{Cov}(\cdot)$ stand for the variance and the covariance operators, respectively. The OLS estimator of $V(\hat{\delta})$ is given by

$$\hat{V}(\hat{\delta}) = \hat{V}(\hat{\alpha}) + \hat{V}(\hat{\beta}) + 2\widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta}).$$

The relevant test-statistic for testing $\delta = 0$ against $\delta \neq 0$ is now given by

$$t_{\hat{\delta}} = \frac{\hat{\delta}}{\sqrt{\hat{V}(\hat{\delta})}} = \frac{\hat{\alpha} + \hat{\beta} - 1}{\sqrt{\hat{V}(\hat{\alpha}) + \hat{V}(\hat{\beta}) + 2\widehat{\mathrm{Cov}}(\hat{\alpha}, \hat{\beta})}}, \tag{3.23}$$

and, under $\delta = 0$, has a t-distribution with $T - 3$ degrees of freedom. An alternative method for testing $\delta = 0$ is the regression method. This starts with (3.21) and replaces β (or α) in terms of δ

and α (or β). Using (3.22) we have

$$\beta = \delta - \alpha + 1.$$

Substituting this in (3.21) for β now yields

$$LY_t - LL_t = a + \alpha(LK_t - LL_t) + \delta LL_t + u_t,$$

or

$$Z_t = a + \alpha W_t + \delta LL_t + u_t, \tag{3.24}$$

where $Z_t = \log(Y_t/L_t) = LY_t - LL_t$ and $W_t = \log(K_t/L_t) = LK_t - LL_t$. A test of $\delta = 0$ can now be carried out by first regressing Z_t on W_t and LL_t (including an intercept term), and then carrying out the usual t-test on the coefficient of LL_t in (3.24). The t-ratio of δ in (3.24) will be identical to $t_{\hat{\delta}}$ defined by (3.23). We now apply the two methods discussed above to the historical data on Y, K, and L used originally by Cobb and Douglas (1928), covering the period 1899–1922. The following estimates of $\hat{\alpha}$, $\hat{\beta}$ and of the variance covariance matrix of $(\hat{\alpha}, \hat{\beta})'$ can be obtained:

$$\hat{\alpha} = 0.23305, \quad \hat{\beta} = 0.80728,$$

$$\begin{bmatrix} \hat{V}(\hat{\alpha}) & \widehat{\mathrm{Cov}}\left(\hat{\alpha}, \hat{\beta}\right) \\ \widehat{\mathrm{Cov}}\left(\hat{\alpha}, \hat{\beta}\right) & \hat{V}(\hat{\beta}) \end{bmatrix} = \begin{bmatrix} 0.004036 & -0.0083831 \\ -0.0083831 & 0.021047 \end{bmatrix}.$$

Using the above results in (3.23) yields

$$t_{\hat{\delta}} = \frac{0.23305 + 0.80728 - 1}{\sqrt{0.004036 + 0.021047 - 2(0.0083831)}} = 0.442. \tag{3.25}$$

Comparing $t_{\hat{\delta}} = 0.442$ and the 5 per cent critical value of the t-distribution with $T - 3 = 24 - 3 = 21$ degrees of freedom (which is equal to 2.080), it is clear that since $t_{\hat{\delta}} = 0.442 < 2.080$, then the hypothesis $\delta = 0$ or $\alpha + \beta = 1$ cannot be rejected at the 5 per cent level. Implementing the regression approach, we estimate (3.24) by OLS and obtain estimates for the coefficients of W_t and LL_t of 0.2330(0.06353) and 0.0403(0.0912), respectively. (The figures in brackets are standard errors.) Note that the t-ratio of the coefficient of the LL variable in this regression is equal to $0.0403/0.0912 = 0.442$, which is identical to $t_{\hat{\delta}}$ as computed in (3.25). It is worth noting that the estimates of α and β, which have played a historically important role in the literature, are very 'fragile', in the sense that they are highly sensitive to the sample period chosen in estimating them. For example, estimating the model (given in (3.21)) over the period 1899–1920 (dropping the observations for the last two years) yields $\hat{\alpha} = 0.0807(0.1099)$ and $\hat{\beta} = 1.0935(0.2241)$.

3.7 Joint tests of linear restrictions

So far we have considered testing a *single* linear restriction on the regression coefficients. Suppose now that we are interested in testing two or more linear restrictions, jointly. One simple example

is the joint test of zero restrictions on the regression coefficients:

$$H_0 : \beta_1 = \beta_2 = 0,$$
$$H_1 : \beta_1 \neq 0 \qquad \text{and/or} \qquad \beta_2 \neq 0.$$

Note that this joint hypothesis is different from testing the following two hypotheses separately.

$$\begin{cases} H_0^I : \beta_1 = 0, \\ H_1^I : \beta_2 \neq 0. \end{cases} \qquad \text{or} \qquad \begin{cases} H_0^{II} : \beta_2 = 0, \\ H_1^{II} : \beta_1 \neq 0. \end{cases}$$

The latter tests are known as separate induced tests and could lead to test outcomes that differ from the outcome of a joint test.

The general procedure for testing joint hypotheses in regression contexts is to construct the *F*-statistic that compares the sum of squares of residuals (SSR) of the regression under the restrictions (i.e., under H_0) with the SSR under the alternative hypothesis, H_1, when the parameter restrictions are not applied. This procedure is valid for a two-sided test. Carrying out one sided tests in the case of joint hypotheses is more complicated and will not be addressed here.

The relevant statistic for the joint test of $r \leq k$ different linear restrictions on the regression coefficients is

$$F = \left(\frac{T-k-1}{r}\right)\left(\frac{SSR_R - SSR_U}{SSR_U}\right), \tag{3.26}$$

where

$SSR_R \equiv$ Restricted sum of squares of errors (residuals)
$SSR_U \equiv$ Unrestricted sum of squares of errors
$k \equiv$ Number of regression coefficients, excluding the intercept term
$T \equiv$ Number of observations
$r \equiv$ Number of independent linear restrictions on the regression coefficients.

Under the null hypothesis, the above statistic, *F*, has an *F*-distribution with *r* and $T - k - 1$ degrees of freedom.

Consider now the application of this general procedure to the problem of testing $\beta_1 = \beta_2 = 0$. The restricted sum of squares of errors (SSR_R) for the problem is obtained by imposing the restrictions $\beta_1 = \beta_2 = 0$ on (3.17) and then by estimating the restricted model

$$y_t = \beta_0 + u_t.$$

This yields $\hat{\beta}_0 = \bar{y}$ and hence

$$SSR_R = \sum_t \left(y_t - \bar{y}\right)^2 = S_{YY}.$$

The unrestricted sum of squares of errors is given by

$$\begin{aligned} SSR_U &= \sum_t \left(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_{t1} - \hat{\beta}_2 x_{t2}\right)^2 \\ &= \sum_t \hat{u}_t^2. \end{aligned}$$

Hence

$$F = \left(\frac{T-3}{2}\right)\left(\frac{S_{YY} - \sum_t \hat{u}_t^2}{\sum_t \hat{u}_t^2}\right),$$

which under the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$, has an F-distribution with 2 and $T-3$ degrees of freedom. The joint hypothesis is rejected if F is larger that the $(1-\alpha)$ per cent critical value of the F-distribution with 2 and $T-3$ degrees of freedom.

3.8 Testing general linear restrictions

All the above tests can be derived as a special case of tests of the following r general linear restrictions

$$\begin{aligned} H_0 &: \mathbf{R}\boldsymbol{\beta} - \mathbf{d}_0 = \mathbf{0}, \\ H_1 &: \mathbf{R}\boldsymbol{\beta} - \mathbf{d}_0 \neq \mathbf{0}, \end{aligned}$$

where $\mathbf{R}$ is an $r \times k$ matrix of known constants with full row rank given by $r \leq k$, and $\mathbf{d}$ is an $r \times 1$ vector of constants. The different hypotheses considered above can be obtained by appropriate choice of $\mathbf{R}$ and $\mathbf{d}_0$. For example, if the object of the exercise is to test the null hypothesis that the first element of $\boldsymbol{\beta}$ is equal to zero, then we need to set $\mathbf{R} = (1, 0, \ldots, 0)$, and $\mathbf{d}_0$=0. To test the hypothesis that the sum of the first two elements adds up to 2 and the sum of the second two elements of $\boldsymbol{\beta}$ adds up to 3 we set

$$\mathbf{R} = \begin{pmatrix} 1 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 1 & 0 & \ldots & 0 \end{pmatrix}, \mathbf{d}_0 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}.$$

The F-statistic for testing H_0 is given by

$$F = \frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{d}_0\right)' \left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{d}_0\right)}{r\hat{\sigma}^2}, \tag{3.27}$$

where $\hat{\boldsymbol{\beta}}$ is the unrestricted *OLS* estimator of $\boldsymbol{\beta}$, and $\hat{\sigma}^2 = (\mathbf{y} - X\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/(T-k)$ is the unbiased estimator of σ^2. Using the distributional results obtained in Chapter 2, in particular the result given by (2.28), it follows that under H_0 the F statistic given by (3.27) has a central F-distribution with r and $T-k$ degrees of freedom. This result of course requires that the classical normal regression assumptions A1–A5 set out in Chapter 2 hold.

3.8.1 Power of the F-test

To obtain the power of the F-test defined by (3.27), consider the alternative hypothesis, H_1, where $\mathbf{R}\boldsymbol{\beta} = \mathbf{d}_1$, and recall that $\mathbf{R}$ is an $r \times k$ matrix of constants with full column rank r. Note that, under H_1,

$$\begin{aligned}\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{d}_0 &= \mathbf{R}\boldsymbol{\beta} - \mathbf{d}_0 + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \boldsymbol{\delta} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \sim N(\boldsymbol{\delta}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'),\end{aligned}$$

where $\boldsymbol{\delta} = \mathbf{d}_1 - \mathbf{d}_0$. Hence

$$X_1 = \frac{\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{d}_0\right)'\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}\left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{d}_0\right)}{\sigma^2} \sim \chi^2_r(\lambda), \tag{3.28}$$

where $\chi^2_r(\lambda)$ is a non-central chi-square variate with r degrees of freedom and the non-centrality parameter[1]

$$\lambda = \frac{\boldsymbol{\delta}'\left[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1}\boldsymbol{\delta}}{\sigma^2} = \boldsymbol{\delta}'\left[\mathbf{R}Var(\hat{\boldsymbol{\beta}})R'\right]^{-1}\boldsymbol{\delta}. \tag{3.29}$$

Furthermore, from (2.27) we know that $X_2 = (T-k)\,\hat{\sigma}^2/\sigma^2 \sim \chi^2_{T-k}$. Using a similar line of reasoning as in Chapter 2, it is easily seen that X_1 (defined by (3.28)) and X_2 are independently distributed, and hence under H_1 the F-statistic given by (3.27) is distributed as a non-central F-distribution with r and $T-k$ degrees of freedom, and the non-centrality parameter, λ, given by (3.29). For given values of r and k, the power of the F test is monotonically increasing in λ. It is clear that the power is higher the greater the distance between the null and the alternative hypotheses as measured by $\boldsymbol{\delta}$, and the greater the precision with which the *OLS* estimators are estimated, as measured by the inverse of $Var(\hat{\boldsymbol{\beta}})$.

3.9 Relationship between the *F*-test and the coefficient of multiple correlation

The relationship between the correlation coefficient and the t-statistic discussed earlier can be readily extended to the multivariate context. Consider the multivariate regression model

$$y_t = \beta_0 + \sum_{j=1}^{k}\beta_j x_{tj} + u_t, \qquad t = 1, 2, \ldots, T,$$

[1] For further information regarding the non-central chi-square distribution see Section B.10.2 in Appendix B.

and suppose we are interested in testing the joint significant of the regressors $x_{t1}, x_{t2}, \ldots, x_{tk}$. The relevant hypothesis is

$$H_0 : \beta_1 = \beta_2, \cdots = \beta_k = 0,$$
$$H_1 : \beta_1 \neq 0, \beta_2 \neq 0, \cdots \beta_k \neq 0.$$

The F-test for this test is given by

$$F = \left(\frac{T-k-1}{k}\right)\left(\frac{S_{YY} - \sum_t \hat{u}_t^2}{\sum_t \hat{u}_t^2}\right),$$

The multiple correlation coefficient is defined by (see (2.30))

$$R^2 = 1 - \frac{\sum_t \hat{u}_t^2}{S_{YY}}.$$

Hence

$$F = \left(\frac{T-k-1}{k}\right)\left(\frac{S_{YY}}{\sum_t \hat{u}_t^2} - 1\right) = \left(\frac{T-k-1}{k}\right)\left(\frac{R^2}{1-R^2}\right),$$

which yields the generalization of the result (3.7) obtained in the case of the simple regression.

3.10 Joint confidence region

To construct a joint confidence region of size $(1-\alpha) \times 100$ for $\left(\beta_1, \beta_2, \ldots, \beta_k\right)$, we first note that the combination of the confidence intervals (3.16) constructed for each β_j separately *does not* yield a joint confidence region with the correct size (namely $1-\alpha$). This is because of dependence of the estimated regression coefficients on each other. Only in the case where $\text{Cov}\left(\hat{\beta}_i, \hat{\beta}_j\right) = 0$ for all $i \neq j$, the joint confidence region of $\left(\beta_1, \beta_2, \ldots, \beta_k\right)$ coincides with the intersection of the confidence intervals obtained for each regression coefficient separately. The appropriate joint confidence region for $\beta_1, \beta_2, \ldots, \beta_k$ is constructed using the F-statistic.

The $(1-\alpha) \times 100$ per cent joint confidence region for β_1 and β_2 in the three variable regression model (2.15) is an ellipsoid in the β_1 and β_2 plane. The shape and the position of this ellipsoid is determined by the size of the confidence region, $1-\alpha$, the *OLS* estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ and the degree of the statistical dependence between the estimators of β_1 and β_2. In matrix notations the formula for this ellipsoid is given by

$$F_\alpha\,(2, T-3) = \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)' \left[\widehat{\text{Cov}}\left(\hat{\boldsymbol{\beta}}\right)\right]^{-1} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right), \tag{3.30}$$

where $\boldsymbol{\beta} = \left(\beta_1, \beta_2\right)'$,

$$\widehat{\text{Cov}}\left(\hat{\boldsymbol{\beta}}\right) = \begin{pmatrix} \widehat{Var}\left(\hat{\beta}_1\right) & \widehat{\text{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) \\ \widehat{\text{Cov}}\left(\hat{\beta}_1, \hat{\beta}_2\right) & \widehat{Var}\left(\hat{\beta}_2\right) \end{pmatrix},$$

and $F_\alpha\ (2, T-3)$ is the $(1-\alpha) \times 100$ per cent critical value of the F-distribution with 2 and $T-3$ degrees of freedom.

3.11 The multicollinearity problem

Multicollinearity is commonly attributed to situations where there is a high degree of intercorrelations among the explanatory variables in a multivariate regression equation. Multicollinearity is particularly prevalent in the case of time series data where there often exists the same common trend in two or more regressors in the regression equation. As a simple example consider the model

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + u_t, \tag{3.31}$$

and assume for simplicity that (x_{t1}, x_{t2}) have a bivariate distribution with the correlation coefficient, ρ_{12}. That is

$$\rho_{12} = \frac{\text{Cov}\ (x_{t1}, x_{t2})}{\left[Var\ (x_{t1})\ Var\ (x_{t2})\right]^{\frac{1}{2}}}.$$

It is clear that as ρ approaches unity *separate* estimation of the slope coefficients β_1 and β_2 becomes more and more problematic. Multicollinearity (namely a value of ρ_{12} near unity in the context of the present example) will be a problem if x_{t1} and x_{t2} are jointly statistically significant but neither is statistically significant when taken individually. Put differently, multicollinearity will be a problem when the hypothesis $\beta_1 = 0$ and $\beta_2 = 0$ can not be rejected when tested separately, while the joint hypothesis that $\beta_1 = \beta_2 = 0$ is rejected. This clearly happens when x_{t1} (or x_{t2}) is an exact linear function of x_{t2} (or x_{t1}). In this case $x_{t2} = \gamma x_{t1}$ and (3.31) reduces to the simple regression equation

$$y_t = \alpha + \left(\beta_1 + \beta_2 \gamma\right) x_{t1} + u_t,$$

and it is only possible to estimate $\beta_1 + \gamma \beta_2$. Neither β_1 nor β_2 can be estimated (or tested) separately. This is the case of 'perfect multicollinearity' and arises out of faulty specification of the regression equation. One important example is when four seasonal dummies are included in a quarterly regression model that already contains an intercept term. In general the multicollinearity problem is likely to arise when ρ_{12}^2 is close to 1.

The multicollinearity problem is also closely related to the problem of low power when testing hypotheses concerning the values of the regression coefficients separately. It is worth noting that no matter how large the correlation coefficient between x_{t1} and x_{t2}, so long as it is not exactly equal to ± 1, a test of $\beta_1 = 0$ (or $\beta_2 = 0$) will have the correct size. The high degree of

correlation between x_{t1} and x_{t2} causes the power of the test to be rather low and as a result we may end up not rejecting the null hypothesis that $\beta_1 = 0$ even if it is false.

Example 7 *To demonstrate the multicollinearity problem and its relation to the problem of low power, using Microfit 5.0 we generated 1,000 observations on x_1, x_2 and y in the following manner.*

$$\begin{aligned} x_1 &\sim N(0,1), \\ x_2 &= x_1 + 0.15v, \\ v &\sim N(0,1), \\ y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \\ u &\sim N(0,1), \end{aligned}$$

with $\beta_0 = \beta_1 = \beta_2 = 1$ and where x_1, v and u were generated as independent standardized normal variates using respectively the 'seeds' of 123, 321 and 4321 in the normal random generator. The Microfit batch file for this exercise is called MULTI.BAT and contains the following instructions:

```
SAMPLE   1   1000;
X1 = NORMAL(123);
V = NORMAL(321);
U = NORMAL(4321);
X2 = X1+0.15*V;
Y = 1 + X1 + X2 + U;
```

Now running the regression of y on x_1 and x_2 (including an intercept term) using only the first fifty observations yields

$$\underset{\quad}{y_t =} \underset{(0.1299)}{0.9047} + \underset{(1.0403)}{1.0950}\, x_{t1} + \underset{(1.0200)}{0.8719}\, x_{t2} + \hat{u}_t \qquad t = 1, 2, \ldots, 50, \tag{3.32}$$

$$R^2 = 0.8498, \qquad \hat{\sigma} = 0.8890, \qquad F_{2,47} = 132.98, \tag{3.33}$$

The standard errors of the parameter estimates are given in brackets, R is the multiple correlation coefficient, $\hat{\sigma}$ is the estimated standard error of the regression equation, and $F_{2,47}$ is the F-statistics for testing the joint hypothesis

$$H_0^J : \beta_1 = \beta_2 = 0,$$

against

$$H_1^J : \beta_1 \neq 0, \quad \beta_2 \neq 0.$$

The t*-statistics for the test of the separate induced tests of*

$$H_0^I : \beta_1 = 0$$

against

$$H_1^I : \beta_1 \neq 0,$$

and of

$$H_0^{II} : \beta_2 = 0,$$

against

$$H_1^{II} : \beta_2 \neq 0,$$

It is firstly clear that since the value of the F*-statistic* ($F_{2,47} = 132.98$) *for the test of* $H_0^J : \beta_1 = \beta_2 = 0$ *is well above the 95 critical value of the* F*-distribution with 2 and 47 degrees of freedom, we conclude that the joint hypothesis* $\beta_1 = \beta_2 = 0$ *is rejected at least at the 95 per cent significance level. Turning now to the tests of* $\beta_1 = 0$ *and* $\beta_2 = 0$ *separately, (i.e. testing the separate induced null hypotheses* H_0^I *and* H_0^{II}*), we note that the* t*-statistics for these hypotheses are equal to* $t_{\hat{\beta}_1} = 1.0950/1.0403 = 1.05$ *and* $t_{\hat{\beta}_2} = 0.8719/1.0200 = 0.85$*, respectively. Neither is statistically significant and the null hypothesis of* $\beta_1 = 0$ *or* $\beta_2 = 0$ *can not be rejected. There is clearly a multicollinearity problem. The joint hypothesis that* β_1 *and* β_2 *are both equal to zero is strongly rejected, but neither of the hypotheses that* β_1 *and* β_2 *are separately equal to zero can be rejected. The sample correlation coefficient of* x_1 *and* x_2 *computed using the first 50 observations is equal to 0.99316 which is apparently too high, given the sample size and the fit of the underlying equation, for the* β_1 *and* β_2 *coefficients to be estimated separately with any degree of precision. In short, the separate induced tests lack the necessary power to allow rejection of* $\beta_1 = 0$ *and* $\beta_2 = 0$ *separately. The relationship between the* F*-statistic used to test the joint hypothesis* $\beta_1 = \beta_2 = 0$*, and the* t*-statistics used to test* $\beta_1 = 0$ *and* $\beta_2 = 0$ *separately, can also be obtained theoretically. Recall from Section 3.7 that*

$$F = \left(\frac{T-3}{2}\right)\left(\frac{S_{YY} - \sum_t \hat{u}_t^2}{\sum_t \hat{u}_t^2}\right). \tag{3.34}$$

Denote the t-statistics for testing $\beta_1 = 0$ *and* $\beta_2 = 0$ *separately by* t_1 *and* t_2*, respectively. Then*

$$t_j^2 = \frac{\hat{\beta}_j^2}{\widehat{Var}\left(\hat{\beta}_j\right)}, \qquad j = 1, 2.$$

But using results in Example 1 (Chapter 2)

$$\widehat{Var}\left(\hat{\beta}_1\right) = \frac{\hat{\sigma}^2 S_{22}}{S_{11}S_{22} - S_{12}^2},$$
$$\widehat{Var}\left(\hat{\beta}_2\right) = \frac{\hat{\sigma}^2 S_{11}}{S_{11}S_{22} - S_{12}^2},$$

where as before $S_{js} = \sum_t (x_{tj} - \bar{x}_j)(x_{ts} - \bar{x}_s)$. *Also since* $y_t - \bar{y} = \hat{\beta}_1 (x_{t1} - \bar{x}_1) + \hat{\beta}_2 (x_{t2} - \bar{x}) + \hat{u}_t$ *we have*[2]

$$S_{YY} = \sum_t (y_t - \bar{y})^2 = \hat{\beta}_1^2 S_{11} + \hat{\beta}_2^2 S_{22} + 2\hat{\beta}_1 \hat{\beta}_2 S_{12} + \sum_t \hat{u}_t^2.$$

Using these results in the expression for the F*-statistic in (3.34) we obtain:*

$$F = \frac{t_1^2 + t_2^2 + 2\rho_{12} t_1 t_2}{2(1 - \rho_{12}^2)}, \tag{3.35}$$

where ρ_{12} *is the sample correlation coefficient between* x_{t1} *and* x_{t2}.[3] *This relationship clearly shows that even for small values of* t_1 *and* t_2 *it is possible to get quite large values of F so long as* ρ_{12} *is chosen to be close enough to 1.*

The above example considers the simple case of a regression model with two explanatory variables. In case of regression models with more than two regressors the detection of the multicollinearity problem becomes more complicated. For example, when there are three regressors with the coefficients β_1, β_2 and β_3, we need to consider all the possible combinations of the coefficients, namely testing them separately: $\beta_1 = 0, \beta_2 = 0, \beta_3 = 0$; in pairs: $\beta_1 = \beta_2 = 0$, $\beta_2 = \beta_3 = 0, \beta_1 = \beta_3 = 0$; and jointly: $\beta_1 = \beta_2 = \beta_3$. Only in the case where the results of separate induced tests, the 'pairs' tests and the joint test are free from contradictions can we be confident that multicollinearity is not a problem.

There exist a number of measures in the literature that purport to detect and measure the seriousness of the multicollinearity problem. One commonly used diagnostic is the *condition number* defined as the square root of the ratio of the largest to the smallest eigenvalue of the matrix $\mathbf{X}'\mathbf{X}$, where the columns of $\mathbf{X}$ have been re-scaled to length 1 (namely, the elements of the j^{th} column of $\mathbf{X}$ have been divided by $s_j = (\sum_{t=1}^{T} x_{tj}^2)^{1/2}$, for $j = 1, 2, \ldots, k$). The condition number detects whether the matrix $\mathbf{X}'\mathbf{X}$ has a small determinant, namely if it is *ill-conditioned*. The larger the condition number, the more ill-conditioned is the matrix, and difficulties can be encountered in calculations involving $(\mathbf{X}'\mathbf{X})^{-1}$. Values of condition number higher than 30 are suggested as indicative of a problem (see Belsley, Kuh, and Welsch (1980) for details). Another diagnostic used to detect multicollinearity is the *variance-inflation factor (VIF)*, defined as $VIF_j = (1 - R_j^2)^{-1}$, for the j^{th} regressor, where R_j^2 is the squared multiple correlation coefficient of the regression of x_{tj} on all other variables in the regression. A high value of VIF_j suggests that x_{tj} is in some collinear relationship with the other regressors. As a rule of thumb, for scaled data, a VIF_j higher than ten indicates severe collinearity (see Kennedy (2003)). We remark that these measures only examine the inter-correlation between the regressors, and at best give a partial picture of the multicollinearity problem, and can often 'lead' to misleading conclusions.

[2] Note that the *OLS* residuals are orthogonal to the regressors.

[3] In the simulation exercise we obtained $t_1 = 1.05$, $t_2 = 0.85$ and $\rho_{12} = 0.99316$. Using these estimates in (3.35) yields $F = 131.50$, which is of the same order of magnitude as the *F*-statistic reported in (3.34). The difference between the two values is due to the error of approximations.

A useful rule of thumb which goes beyond regressor correlations is to compare the squared multiple correlation coefficient of the regression equation, R^2, with R_j^2. Klein (1962) suggests that collinearity is likely to be a problem and could lead to imprecise estimates if $R^2 < R_j^2$, for some $j = 1, 2, \ldots, k$.

Example 8 *To illustrate the problem return to the simulation exercise, and use the first 500 observations (instead of the first 50 observations) in computing the regression of y on x_1 and x_2. The results are*

$$\underset{}{y_t} = \underset{(0.0428)}{0.9307} + \underset{(0.28343)}{1.1045}\, x_{t1} + \underset{(0.27081)}{0.93138}\, x_{t2} + \hat{u}_t \qquad t = 1, 2, \ldots, 500,$$

$$R^2 = 0.8333, \qquad \hat{\sigma} = 0.95664, \qquad F_{2,497} = 1242.3.$$

As compared with the estimates based on the first 50 observations [see (3.32) and (3.33)], these estimates have much smaller standard errors and using the 95 percent significance level we arrive at similar conclusions whether we test $\beta_1 = 0$ and $\beta_2 = 0$ separately or jointly. Yet the sample correlation coefficient between x_{t1} and x_{t2} estimated over the first 500 observations is equal to 0.9895 which is only marginally smaller than the estimate obtained for the first 50 observations. By increasing the sample size from 50 to 500 we have increased the precision with which β_1 and β_2 are estimated and the power of testing $\beta_1 = 0$ and $\beta_2 = 0$ both separately and jointly.

The above illustration also points to the fact that the main cause of the multicollinearity problem is lack of adequate observations (or information), and hence the imprecision with which the parameters of interest are estimated. Assuming the regression model under consideration is correctly specified, the only valid solution to the problem is to increase the information on the basis of which the regression is estimated. The new information could be either in the form of additional observations on y, x_1 and x_2, or it could be some *a priori* information concerning the parameters. The latter fits well with the Bayesian approach, but is difficult to accommodate within the classical framework. There are also other approaches suggested in the literature such as the ridge regression, and the principle component regression to deal with the multicollinearity problem. For a Bayesian treatment of the regression analysis see Section C.6 in Appendix C. However, in using Bayesian techniques to deal with the multicollinearity problem it is important to bear in mind that the posterior means of the regression coefficients are well defined in small samples even if the regressors are highly multicollinear and even if $\mathbf{X}'\mathbf{X}$ is rank deficient. But in such cases the posterior mean of $\boldsymbol{\beta}$ can be very sensitive to the choice of the priors, and unless $T^{-1}\mathbf{X}'\mathbf{X}$ tends to a positive definite matrix the Bayes estimates of $\boldsymbol{\beta}$ could become unstable as $T \to \infty$.

Example 9 *As an example consider the following Fisher type explanation of nominal interests estimated on US quarterly data over the period 1948(1)–1990(4) using the file USGNP.FIT provided in Microfit 5:*

$$\begin{aligned} R_t = {} & \underset{(0.1295)}{-0.0381} + \underset{(0.0754)}{1.2606}\, R_{t-1} \underset{(0.1144)}{-.61573}\, R_{t-2} + \underset{(0.1208)}{0.6073}\, R_{t-3} - \\ & \underset{(0.0782)}{0.3168}\, R_{t-4} + \underset{(0.1075)}{0.13198}\, DM_{t-1} + \underset{(0.1064)}{0.1072}\, DM_{t-2} + \hat{u}_t, \end{aligned}$$

$$R^2 = 0.9520, \qquad \bar{R}^2 = 0.9503, \qquad \hat{\sigma} = 0.7086, \qquad F_{6,165} = 545.83,$$

where R_t = nominal rate of interest, DM_t = the growth of money supply (M_2 definition). In this regression, the coefficients of the lagged interest rate variables are all significant, but neither of the two coefficients of the lagged monetary growth variable is statistically significant. The t-ratios for the coefficients of DM_{t-1} and DM_{t-2} are equal to 1.23 and 1.01, respectively, while the 95 percent critical value of the t-distribution with 165 (namely $T - k = 172 - 7$) degrees of freedom is equal to 1.97. As we have seen above, it would be a mistake to necessarily conclude from this result that monetary growth has no significant impact on the nominal interest rates in the US. The statistical insignificance of the coefficients of DM_{t-1} and DM_{t-2}, when tested separately may be due to the high intercorrelation between the regressors. Also we are not interested in testing the statistical significance of individual coefficients of the past monetary growth rates. What is of interest is the sum of the two coefficients of the lagged monetary growth rates, and not the individual coefficients, separately. Denote the coefficients of DM_{t-1} and DM_{t-2} by γ_1 and γ_2 respectively, and let $\delta = \gamma_1 + \gamma_2$. We have

$$\hat{\delta} = \hat{\gamma}_1 + \hat{\gamma}_2 = 0.1319 + 0.1072 = 0.2391$$

To compute the estimate of the standard error of $\hat{\delta}$ we recall that

$$\widehat{Var}\left(\hat{\delta}\right) = \widehat{Var}\left(\hat{\gamma}_1\right) + \widehat{Var}\left(\hat{\gamma}_2\right) + 2\widehat{\mathrm{Cov}}\left(\hat{\gamma}_1, \hat{\gamma}_2\right),$$

and using the Microfit package we have

$$\widehat{Var}\left(\hat{\gamma}_1\right) = 0.01156, \qquad \widehat{Var}\left(\hat{\gamma}_2\right) = 0.01132, \qquad \widehat{\mathrm{Cov}}\left(\hat{\gamma}_1, \hat{\gamma}_2\right) = -0.00854$$

and hence $\sqrt{\widehat{Var}\left(\hat{\delta}\right)} = 0.0762$, and $t_\delta = 0.2391/0.0762 = 3.14$ which is well above the 95 percent critical value of the t-distribution with 165 degrees of freedom. Therefore, we strongly reject the hypothesis that monetary growth has no effect on the nominal interest in the US. We also note that for every one percent increase in the growth of money supply there is around 0.24 of one percent increase in nominal interest within the space of two quarters. The long-run impact of money supply growth on nominal interest is much larger and depends on the magnitude of the lagged coefficients of the nominal interest rates.

3.12 Multicollinearity and the prediction problem

Consider the following regression model

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u},$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_T)'$, $\mathbf{X}_1$ and $\mathbf{X}_2$ are $T \times k_1$ and $T \times k_2$ regressor matrices that are perfectly correlated, namely

$$\mathbf{X}_2 = \mathbf{X}_1\mathbf{A}',$$

and $\mathbf{A}$ is a $k_2 \times k_1$ matrix of fixed constants. Further assume that $\mathbf{X}_1'\mathbf{X}_1$ is a positive definite matrix. Consider now the forecast of y_{T+1} conditional on $\mathbf{x}_{T+1} = (\mathbf{x}_{1T}', \mathbf{x}_{2T}')'$ which is given by[4]

$$\hat{y}_{T+1} = \mathbf{x}_{T+1}'\hat{\boldsymbol{\beta}}_T = \mathbf{x}_{T+1}'(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{y},$$

where $(\mathbf{X}'\mathbf{X})^{+}$ is the generalized inverse of $\mathbf{X}'\mathbf{X}$, defined by (see also Section A.7)

$$\left(\mathbf{X}'\mathbf{X}\right)(\mathbf{X}'\mathbf{X})^{+}\left(\mathbf{X}'\mathbf{X}\right) = \left(\mathbf{X}'\mathbf{X}\right).$$

It is well known that $(\mathbf{X}'\mathbf{X})^{+}$ is not unique when $\mathbf{X}'\mathbf{X}$ is rank deficient. In what follows we show that $\hat{y}_{T+1}$ is unique despite the non-uniqueness of $(\mathbf{X}'\mathbf{X})^{+}$. Note that

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_1\mathbf{A}' \\ \mathbf{A}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{A}\mathbf{X}_1'\mathbf{X}_1\mathbf{A}' \end{pmatrix} = \mathbf{H}\mathbf{X}_1'\mathbf{X}_1\mathbf{H}',$$

where $\mathbf{H}$ is a $k \times k_1$ matrix $(k = k_1 + k_2)$:

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_{k_1} \\ \mathbf{A} \end{pmatrix}.$$

Also

$$\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{X}_1'\mathbf{y}, \text{ and } \mathbf{x}_{T+1}' = \mathbf{x}_{1T}'\mathbf{H}'.$$

Hence

$$\hat{y}_{T+1} = \mathbf{x}_{T+1}'(\mathbf{X}'\mathbf{X})^{+}\mathbf{X}'\mathbf{y} = \mathbf{x}_{1T}'\mathbf{H}'\left(\mathbf{H}\mathbf{X}_1'\mathbf{X}_1\mathbf{H}'\right)^{+}\mathbf{H}\mathbf{X}_1'\mathbf{y}.$$

Since $\mathbf{X}_1'\mathbf{X}_1$ is a symmetric positive definite matrix, then

$$\hat{y}_{T+1} = \mathbf{x}_{1T}'\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1/2}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{1/2}\mathbf{H}'\left(\mathbf{H}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{1/2}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{1/2}\mathbf{H}'\right)^{+}$$
$$\mathbf{H}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{1/2}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1/2}\mathbf{X}_1'\mathbf{y},$$

or

$$\hat{y}_{T+1} = \mathbf{x}_{1T}'\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1/2}\mathbf{G}'\left(\mathbf{G}\mathbf{G}'\right)^{+}\mathbf{G}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1/2}\mathbf{X}_1'\mathbf{y},$$

where

$$\mathbf{G} = \mathbf{H}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{1/2}.$$

Consider now the $k_1 \times k_1$ matrix $\mathbf{G}\left(\mathbf{G}'\mathbf{G}\right)^{+}\mathbf{G}'$ and note that from properties of generalized inverse we have

[4] A general treatment of the prediction problem is given in Chapter 17.

$$\left(\mathbf{GG}'\right)\left(\mathbf{GG}'\right)^{+}\left(\mathbf{GG}'\right)=\left(\mathbf{GG}'\right).$$

Pre- and post-multiplying the above by $\mathbf{G}'$ and $\mathbf{G}$, we have

$$\left(\mathbf{G}'\mathbf{G}\right)\mathbf{G}'\left(\mathbf{GG}'\right)^{+}\mathbf{G}\left(\mathbf{G}'\mathbf{G}\right)=\left(\mathbf{G}'\mathbf{G}\right)\left(\mathbf{G}'\mathbf{G}\right). \tag{3.36}$$

But

$$\mathbf{G}'\mathbf{G}=\left(\mathbf{X}_1'\mathbf{X}_1\right)^{1/2}\mathbf{H}'\mathbf{H}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{1/2},$$

and since

$$\mathbf{H}'\mathbf{H}=\mathbf{I}_{k_1}+\mathbf{A}'\mathbf{A},$$

then

$$\mathbf{G}'\mathbf{G}=\left(\mathbf{X}_1'\mathbf{X}_1\right)^{1/2}\left(\mathbf{I}_{k_1}+\mathbf{A}'\mathbf{A}\right)\left(\mathbf{X}_1'\mathbf{X}_1\right)^{1/2},$$

is a nonsingular matrix (for any $\mathbf{A}$) and has a unique inverse. Using this result in (3.36) it now follows that

$$\mathbf{G}'\left(\mathbf{GG}'\right)^{+}\mathbf{G}=\mathbf{I}_{k_1},$$

and hence

$$\begin{aligned}\hat{y}_{T+1}&=\mathbf{x}_{1T}'\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1/2}\mathbf{G}'\left(\mathbf{GG}'\right)^{+}\mathbf{G}\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1/2}\mathbf{X}_1'\mathbf{y}\\&=\mathbf{x}_{1T}'\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{y},\end{aligned}$$

which is unique and invariant to the choice of the generalized inverse of $\mathbf{X}'\mathbf{X}$.

3.13 Implications of misspecification of the regression model on hypothesis testing

Suppose that y_t is generated according to the classical linear regression equation

$$y_t=\beta_0+\beta_1x_t+\beta_2z_t+u_t, \tag{3.37}$$

but the investigator estimates the simple regression equation

$$y_t=\alpha+\beta x_t+\varepsilon_t, \tag{3.38}$$

which omits the regressor z_t. We have seen in Section 2.13 that omitting a relevant regressor, z_t, may lead to biased estimates, unless the included regressor, x_t, and the omitted variable, z_t, are uncorrelated. However, even in the case x_t and z_t are uncorrelated, $\hat{\beta}$ will not be an efficient

estimator of β_1. This is because the correct estimator of the variance of $\hat{\beta}$ requires knowledge of an estimator of $\sigma_u^2 = Var(u_t)$, namely

$$\hat{\sigma}_u^2 = \frac{\sum_t \hat{u}_t^2}{T-3} = \frac{\sum_t \left(y_t - \hat{\beta}_0 - \hat{\beta}_1 x_t - \hat{\beta}_2 z_t\right)^2}{T-3}.$$

with $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$ being *OLS* estimators of parameters in (3.37), while the regression with the omitted variable only yields an estimator of $\sigma_\varepsilon^2 = Var(\varepsilon_t)$, namely

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_t \hat{\varepsilon}_t^2}{T-2} = \frac{\sum_t \left(y_t - \hat{\alpha} - \hat{\beta} x_t\right)^2}{T-2},$$

with $\hat{\alpha}$ and $\hat{\beta}$ being *OLS* estimators of parameters in (3.38). Notice that, in general, $\hat{\sigma}_\varepsilon^2 \geq \hat{\sigma}_u^2$, and therefore the variance of $\hat{\beta}$ will be generally larger than the variance of $\hat{\beta}_1$. A similar problem in the estimation of the variance of estimated regression parameters arises when additional irrelevant variables are included in the regression equation.

3.14 Jarque–Bera's test of the normality of regression residuals

In many applications, particularly involving financial time series, it is important to investigate the extent to which regression errors exhibit departures from normality. There are two important ways that error distributions could deviate from normality: skewness and Kurtosis (or tail-fatness)

$$\begin{aligned} \text{Skewness} &= \sqrt{b_1} = m_3/m_2^{3/2}, \\ \text{Kurtosis} &= b_2 = m_4/m_2^2, \end{aligned}$$

where

$$m_j = \frac{\sum_{t=1}^{T} \hat{u}_t^j}{T}, \qquad j = 1, 2, 3, 4,$$

For a normal distribution $\sqrt{b_1} \approx 0$, and $b_2 \approx 3$. The Jarque–Bera's test of the departures from normality is given by (see Jarque and Bera (1980) and Bera and Jarque (1987))

$$\chi_T^2(2) = T\left\{\tfrac{1}{6} b_1 + \tfrac{1}{24}(b_2 - 3)^2\right\},$$

if the regression contains an intercept term (note that in that case $m_1 = 0$). When the regression does not contain an intercept term, then $m_1 \neq 0$, and the test statistic has the additional term

$$Tb_0 = T\left\{3m_1^2/(2m_2) - m_3 m_1/m_2^2\right\},$$

namely

$$\chi^2_T(2) = T\left\{b_0 + \tfrac{1}{6}b_1 + \tfrac{1}{24}(b_2 - 3)^2\right\}.$$

3.15 Predictive failure test

Consider the following linear regression models specified for each of the two sample periods

$$\mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}_1; \qquad \mathbf{u}_1 \sim N(0, \sigma_1^2\mathbf{I}_{T1}), \tag{3.39}$$

$$\mathbf{y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}_2; \qquad \mathbf{u}_2 \sim N(0, \sigma_2^2\mathbf{I}_{T2}), \tag{3.40}$$

where $\mathbf{y}_r, \mathbf{X}_r, r = 1, 2$, are $T_r \times 1$ and $T_r \times k$ observation matrices on the dependent variable and the regressors for the two sample periods, and $\mathbf{I}_{T_1}$ and $\mathbf{I}_{T_2}$ are identity matrices of order T_1 and T_2, respectively. Combining (3.39) and (3.40) by stacking the observations on the two sample periods now yields

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0}_{T_1 \times T_2} \\ \mathbf{X}_2 & \mathbf{I}_{T2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\delta} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}.$$

The above system of equations may also be written more compactly as

$$\mathbf{y}_0 = \mathbf{X}_0\boldsymbol{\beta}_1 + \mathbf{S}_2\boldsymbol{\delta} + \mathbf{u}_0, \tag{3.41}$$

where $\mathbf{y}_0 = (\mathbf{y}_1', \mathbf{y}_2')', \mathbf{X}_0 = (\mathbf{X}_1', \mathbf{X}_2')'$, and $\mathbf{S}_2$ represents the $(T_1 + T_2) \times T_2$ matrix of T_2 dummy variables, one dummy variable for each observation in the second period. For example, for observation $T_1 + 1$, the first column of $\mathbf{S}_2$ will have unity on its $(T_1 + 1)^{th}$ element and zeros elsewhere. The predictive failure test can now be carried out by testing the hypothesis of $\boldsymbol{\delta} = 0$ against $\boldsymbol{\delta} \neq 0$ in (3.41). This yields the following F-statistic

$$F_{PF} = \frac{(\hat{\mathbf{u}}_0'\hat{\mathbf{u}}_0 - \hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1)/T_2}{\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1/(T_1 - k)} \sim F(T_2, T_1 - k), \tag{3.42}$$

where

- $\hat{\mathbf{u}}_0$ is the *OLS* residual vector of the regression of $\mathbf{y}_0$ on $\mathbf{X}_0$ (i.e., based on the first and the second sample periods together).
- $\hat{\mathbf{u}}_1$ is the *OLS* residual vector of the regression of $\mathbf{y}_1$ on $\mathbf{X}_1$ (i.e., based on the first sample period).

Under the classical normal assumptions, the predictive failure test statistic, F_{PF}, has an exact F-distribution with T_2 and $T_1 - k$ degrees of freedom.

The *LM* version of the above statistic is computed as

$$\chi^2_{PF} = T_2 F_{PF} \overset{a}{\sim} \chi^2(T_2), \tag{3.43}$$

which is distributed as a chi-squared with T_2 degrees of freedom for large T_1 (see Chow (1960), Salkever (1976), Dufour (1980), and Pesaran, Smith, and Yeo (1985), section III.)

It is also possible to test if the predictive failure is due to particular time period(s) by applying the *t*- or the *F*-tests to one or more elements of $\boldsymbol{\delta}$ in (3.41).

3.16 A test of the stability of the regression coefficients: the Chow test

This test is proposed by Chow (1960) and aims at testing the hypothesis that in (3.39) and (3.40) $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, conditional on equality of variances, that is, $\sigma_1^2 = \sigma_2^2$. In econometrics literature this is known as the Chow test, and is known as the analysis of covariance test in the statistics literature (see Scheffe (1959)). The *F*-version of the Chow test statistic is defined by

$$F_{SS} = \frac{(\hat{\mathbf{u}}_0'\hat{\mathbf{u}}_0 - \hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1 - \hat{\mathbf{u}}_2'\hat{\mathbf{u}}_2)/k}{(\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1 + \hat{\mathbf{u}}_2'\hat{\mathbf{u}}_2)/(T_1 + T_2 - 2k)} \sim F(k, T_1 + T_2 - 2k), \tag{3.44}$$

where

- $\hat{\mathbf{u}}_0$ is the *OLS* residual vector for the first two sample periods together
- $\hat{\mathbf{u}}_1$ is the *OLS* residual vector for the first sample period
- $\hat{\mathbf{u}}_2$ is the *OLS* residual vector for the second sample period.

The *LM* version of this test statistic is computed as

$$\chi_{SS}^2 = kF_{SS} \overset{a}{\sim} \chi^2(k). \tag{3.45}$$

For more details see, for example, Pesaran, Smith, and Yeo (1985, p. 285).

3.17 Non-parametric estimation of the density function

Suppose $f(y)$ denotes the density function of a variable Y at point y, and $y_1, y_2, \ldots, y_T$ are observations drawn from $f(.)$. Two general approaches have been proposed to estimate $f(.)$. The first is a parametric method, which assumes that the form for $f(.)$ is known (e.g., normal), except for the few parameters that need to be estimated consistently from data (e.g., the mean and variance). In contrast, the non-parametric approach tries to estimate $f(.)$ directly, without strong assumptions on its form. One simple example of such an estimator is the histogram, although it has the drawback of being discontinuous, and not applicable for estimating the distribution of two or more variables. The non-parametric density estimator takes the following general form

$$\hat{f}(y) = \frac{1}{T}\sum_{t=1}^{T} \frac{1}{h_T} K\left(\frac{y - y_t}{h_T}\right),$$

where $K(\cdot)$ is called kernel function, and h_T is the window width, also called the *smoothing* parameter or *bandwidth*. The kernel function needs to satisfy some regularity conditions typical of probability density functions, for example, $K(-\infty) = K(\infty) = 0$, and $\int_{-\infty}^{+\infty} K(x)\, dx = 1$. There exists a vast literature on the choice of this function. One popular choice is the Gaussian kernel, namely

$$K(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}.$$

Another common choice is the *Epanechnikov* kernel

$$K(y) = \begin{cases} \frac{3}{4}\left(1 - \frac{1}{5}y^2\right)/\sqrt{5}, & \text{if } |y| < \sqrt{5}, \\ 0, & \text{otherwise.} \end{cases}$$

As also pointed by Pagan and Ullah (1999), the choice of K is not critical to the analysis, and the optimal kernel in most cases will only yield modest improvements in the performance of $\hat{f}(y)$, over selections such as the Gaussian kernel.

When implementing density estimates, the choice of the window width, h_T, plays an essential role. One crude way of choosing h_T is by a trial-and-error approach, consisting of looking at several different plots of $\hat{f}(y)$ against y, when $\hat{f}(y)$ is computed for different values of h_T. Other more objective and automatic methods for selecting h_T have been proposed in the literature. One popular choice is the *Silverman rule of thumb*, according to which

$$h_{srot} = 0.9 \cdot A \cdot T^{-\frac{1}{5}}, \tag{3.46}$$

where $A = \min(\sigma, R/1.34)$, σ is the standard deviation of the variable y, R is the interquartile range, and T is the number of observations, see Silverman (1986, p. 47). Another very popular method is the *least squares cross-validation* method, according to which the window width is the value, h_{lscv}, that minimizes the following criterion

$$ISE(h_T) = \frac{1}{T^2 h_T} \sum_{t \neq s}^{T} K_2\left(\frac{y_t - y_s}{h_T}\right) - \frac{2}{T} \sum_{t=1}^{T} \hat{f}_{-t}(y_t), \tag{3.47}$$

where $K_2(.)$ is the convolution of the kernel with itself, defined by

$$K_2(y) = \int_{-\infty}^{+\infty} K(t)\, K(y - t)\, dt.$$

and $\hat{f}_{-t}(y_t)$ is the density estimator obtained after omitting the t^{th} observation. We have

$$\frac{1}{T} \sum_{t=1}^{T} \hat{f}_{-t}(y_t) = \frac{1}{T(T-1)h_T} \sum_{t \neq j}^{T} K\left(\frac{y_t - y_j}{h_T}\right).$$

If K is the Gaussian kernel, then K_2 is $N(0, 2)$, or

$$K_2(y) = (4\pi)^{-1/2} e^{-y^2/4},$$

for the Epanechnikov kernel we have

$$K_2(y) = \begin{cases} \frac{3\sqrt{5}}{100}\left(4 - y^2\right), & \text{if } |y| < \sqrt{5} \\ 0, & \text{otherwise.} \end{cases}.$$

For the Gaussian kernel the expression for $ISE(h_T)$ simplifies to (see Bowman and Azzalini (1997, p. 37))

$$ISE(h_T) = \frac{1}{(T-1)}\phi\left(0, \sqrt{2}h_T\right) + \frac{T-2}{T(T-1)^2}\sum_{t\neq j}^{T}\phi\left(y_t - y_j, \sqrt{2}h_T\right) - \frac{2}{T(T-1)}\sum_{t\neq j}^{T}\phi\left(y_t - y_j, h_T\right),$$

where $\phi(y, \sigma)$ denotes the normal density function with mean 0 and standard deviation σ:

$$\phi(y, \sigma) = (2\pi\sigma^2)^{-1/2}\exp\left(\frac{-y^2}{2\sigma^2}\right).$$

In cases where local minima are encountered we select the bandwidth that corresponds to the local minimum with the largest value for h_T. See Bowman and Azzalini (1997, pp. 33–4). See also Pagan and Ullah (1999), Silverman (1986), Jones, Marron, and Sheather (1996), and Sheather (2004) for further details.

3.18 Further reading

Further material on statistical inference and its application to econometrics can be found in Rao (1973) and Bierens (2005). See also Appendix B for a review of key concepts from probability theory and statistics useful for this chapter. For what concerns non-parametric density estimators, further discussion can be found in Horowitz (2009), which contains a treatment of non-parametric methods within econometrics.

3.19 Exercises

1. Consider the model

$$\log Y_t = \beta_0 + \beta_1 \log L_t + \beta_2 \log K_t + u_t,$$