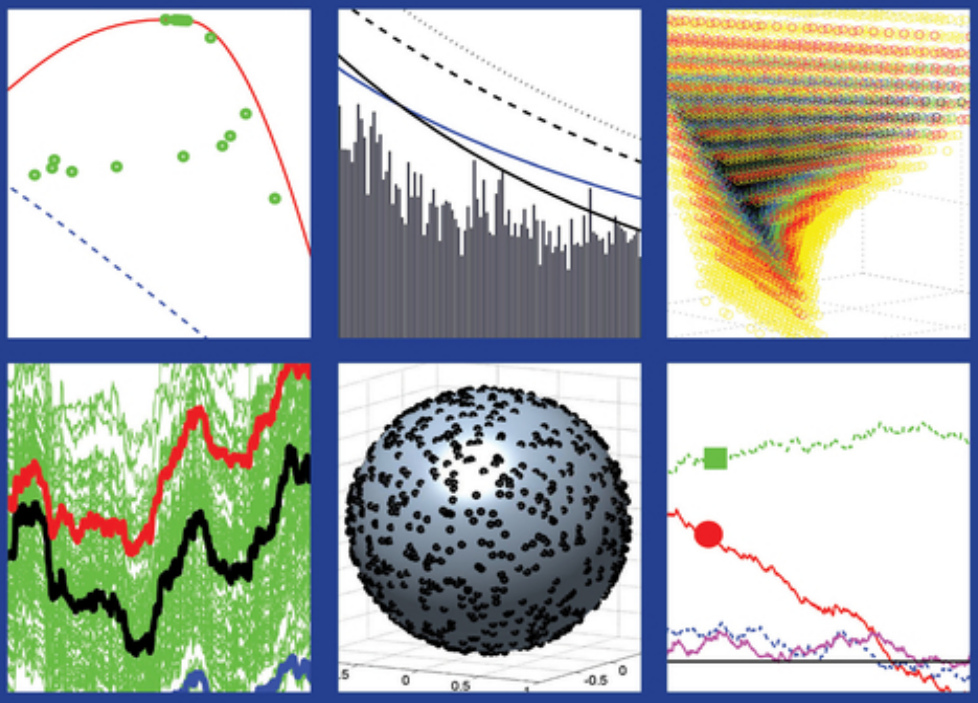
WILEY SERIES IN PROBABILITY AND STATISTICS

Linear Models and Time-Series Analysis

Regression, ANOVA, ARMA and GARCH

Marc S. Paolella



WILEY

Linear Models and Time-Series Analysis

The Wiley Series in Probability and Statistics is well established and authoritative. It covers many topics of current research interest in both pure and applied statistics and probability theory. Written by leading statisticians and institutions, the titles span both state-of-the-art developments in the field and classical methods.

Reflecting the wide range of current research in statistics, the series encompasses applied, methodological and theoretical statistics, ranging from applications and new techniques made possible by advances in computerized practice to rigorous treatment of theoretical approaches.

This series provides essential and invaluable reading for all statisticians, whether in academia, industry, government, or research.

Related Titles

Quantile Regression: Estimation and Simulation, Volume 2 by Marilena Furno, Domenico Vistocco

Nonparametric Finance by Jussi Klemela February 2018

Machine Learning: Topics and Techniques by Steven W. Knox February 2018

Measuring Agreement: Models, Methods, and Applications by Pankaj K. Choudhary, Haikady N. Nagaraja November 2017

Engineering Biostatistics: An Introduction using MATLAB and WinBUGS by Brani Vidakovic October 2017

Fundamentals of Queueing Theory, 5th Edition by John F. Shortle, James M. Thompson, Donald Gross, Carl M. Harris October 2017

Reinsurance: Actuarial and Statistical Aspects by Hansjoerg Albrecher, Jan Beirlant, Jozef L. Teugels September 2017

Clinical Trials: A Methodologic Perspective, 3rd Edition by Steven Piantadosi August 2017

Advanced Analysis of Variance by Chihiro Hirotsu August 2017

Matrix Algebra Useful for Statistics, 2nd Edition by Shayle R. Searle, Andre I. Khuri April 2017

Statistical Intervals: A Guide for Practitioners and Researchers, 2nd Edition by William Q. Meeker, Gerald J. Hahn, Luis A. Escobar March 2017

Time Series Analysis: Nonstationary and Noninvertible Distribution Theory, 2nd Edition by Katsuto Tanaka March 2017

Probability and Conditional Expectation: Fundamentals for the Empirical Sciences by Rolf Steyer, Werner Nagel March 2017

Theory of Probability: A critical introductory treatment by Bruno de Finetti February 2017

Simulation and the Monte Carlo Method, 3rd Edition by Reuven Y. Rubinstein, Dirk P. Kroese October 2016

Linear Models, 2nd Edition by Shayle R. Searle, Marvin H. J. Gruber October 2016

Robust Correlation: Theory and Applications by Georgy L. Shevlyakov, Hannu Oja August 2016

Statistical Shape Analysis: With Applications in R, 2nd Edition by Ian L. Dryden, Kanti V. Mardia July 2016

Matrix Analysis for Statistics, 3rd Edition by James R. Schott June 2016

Statistics and Causality: Methods for Applied Empirical Research by Wolfgang Wiedermann (Editor), Alexander von Eye (Editor) May 2016

Time Series Analysis by Wilfredo Palma February 2016

Linear Models and Time-Series Analysis

Regression, ANOVA, ARMA and GARCH

Marc S. Paolella
Department of Banking and Finance
University of Zurich
Switzerland

WILEY

This edition first published 2019

Registered Offices
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial Office
9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Paolella, Marc S., author.
Title: Linear models and time-series analysis : regression, ANOVA, ARMA and GARCH / Dr. Marc S. Paolella.
Description: Hoboken, NJ : John Wiley & Sons, 2019. | Series: Wiley series in probability and statistics |
Identifiers: LCCN 2018023718 (print) | LCCN 2018032640 (ebook) | ISBN 9781119431855 (Adobe PDF) | ISBN 9781119431985 (ePub) | ISBN 9781119431909 (hardcover)
Subjects: LCSH: Time-series analysis. | Linear models (Statistics)
Classification: LCC QA280 (ebook) | LCC QA280 .P373 2018 (print) | DDC 515.5/5–dc23
LC record available at https://lccn.loc.gov/2018023718

Cover Design: Wiley
Cover Images: Images courtesy of Marc S. Paolella

Set in 10/12pt WarnockPro by SPi Global, Chennai, India

10 9 8 7 6 5 4 3 2 1

Contents

Preface

Cowards die many times before their deaths. The valiant never taste of death but once.
(William Shakespeare, Julius Caesar, Act II, Sc. 2)

The goal of this book project is to set a strong foundation, in terms of (usually small-sample) distribution theory, for the linear model (regression and ANOVA), univariate time-series analysis (ARMAX and GARCH), and some multivariate models associated primarily with modeling financial asset returns (copula-based structures and the discrete mixed normal and Laplace). The primary target audiences of this book are masters and beginning doctoral students in statistics, quantitative finance, and economics.

This book builds on the author's "Fundamental Statistical Inference: A Computational Approach", introducing the major concepts underlying statistical inference in the i.i.d. setting, and thus serves as an ideal prerequisite for this book. I hereafter denote it as book III, and likewise refer to my books on probability theory, Paolella (2006, 2007), as books I and II, respectively. For example, Listing III.4.7 refers to the Matlab code in Program Listing 4.7, chapter 4 of book III, and likewise for references to equations, examples, and pages.

As the emphasis herein is on relatively rigorous underlying distribution theory associated with a handful of core topics, as opposed to being a sweeping monograph on linear models and time series, I believe the book serves as a solid and highly useful prerequisite to larger-scope works. These include (and are highly recommended by the author), for time-series analysis, Priestley (1981), Brockwell and Davis (1991), Hamilton (1994), and Pollock (1999); for econometrics, Hayashi (2000), Pesaran (2015), and Greene (2017); for multivariate time-series analysis, Lütkepohl (2005) and Tsay (2014); for panel data methods, Wooldridge (2010), Baltagi (2013), and Pesaran (2015); for micro-econometrics, Cameron and Trivedi (2005); and, last but far from least, for quantitative risk management, McNeil et al. (2015). With respect to the linear model, numerous excellent books dedicated to the topic are mentioned below and throughout Part I.

Notably in statistics, but also in other quantitative fields that rely on statistical methodology, I believe this book serves as a strong foundation for subsequent courses in (besides more advanced courses in linear models and time-series analysis) multivariate statistical analysis, machine learning, modern inferential methods (such as those discussed in Efron and Hastie (2016), which I mention below), and also Bayesian statistical methods. As also stated in the preface to book III, the latter topic gets essentially no treatment there or in this book, the reasons being (i) to do the subject justice would require a substantial increase in the size of these already lengthy books and (ii) numerous excellent books dedicated to the Bayesian approach, in both statistics and econometrics, and at

varying levels of sophistication, already exist. I believe a strong foundation in underlying distribution theory, likelihood-based inference, and prowess in computing are necessary prerequisites to appreciate Bayesian inferential methods.

The preface to book III contains a detailed discussion of my views on teaching, textbook presentation style, inclusion (or lack thereof) of end-of-chapter exercises, and the importance of computer programming literacy, all of which are applicable here and thus need not be repeated. Also, this book, like books I, II, and III, contains far more material than could be covered in a one-semester course.

This book can be nicely segmented into its three parts, with Part I (and Appendices A and B) addressing the linear (Gaussian) model and ANOVA, Part II detailing the ARMA and ARMAX univariate time-series paradigms (along with unit root testing and time-varying parameter regression models), and Part III dedicated to modern topics in (univariate and multivariate) financial time-series analysis, risk forecasting, and portfolio optimization. Noteworthy also is Appendix C on some multivariate distributional results, with Section C.1 dedicated to the characteristic function of the (univariate and multivariate) Student's t distribution, and Section C.2 providing a rather detailed discussion of, and derivation of major results associated with, the class of elliptic distributions.

A perusal of the table of contents serves to illustrate the many topics covered, and I forgo a detailed discussion of the contents of each chapter.

I now list some ways of (academically) using the book.[1] All suggested courses assume a strong command of calculus and probability theory at the level of book I, linear and matrix algebra, as well as the basics of moment generating and characteristic functions (Chapters 1 and 2 from book II). All courses *except the first* further assume a command of basic statistical inference at the level of book III. Measure theory and an understanding of the Lebesgue integral are *not* required for this book.

In what follows, "Core" refers to the core chapters recommended from this book, "Add" refers to additional chapters from this book to consider, and sometimes other books, depending on interest and course focus, and "Outside" refers to recommended sources to supplement the material herein with important, omitted topics.

1) One-semester beginning graduate course: Introduction to Statistics and Linear Models.
   - Core (not this book):
     Chapters 3, 5, and 10 from book II (multivariate normal, saddlepoint approximations, noncentral distributions).
     Chapters 1, 2, 3 (and parts of 7 and 8) from book III.
   - Core (this book):
     Chapters 1, 2, and 3, and Appendix A.
   - Add: Appendix D.
2) One-semester course: Linear Models.
   - Core (not this book):
     Chapters 3, 5, and 10 from book II (multivariate normal, saddlepoint approximations, noncentral distributions).
   - Core (this book):
     Chapters 1, 2, and 3, and Appendix A.
   - Add: Chapters 4 and 5, and Appendices B and D, select chapters from Efron and Hastie (2016).

1 Thanks to some creative students, other uses of the book include, besides a door stop and useless coffee-table centerpiece, a source of paper for lining the bottom of a bird cage and for mopping up oil spills in the garage.

- Outside (for regression): Select chapters from Chatterjee and Hadi (2012), Graybill and Iyer (1994), Harrell, Jr. (2015), Montgomery et al. (2012).[2]
- Outside (for ANOVA and mixed models): Select chapters from Galwey (2014), West et al. (2015), Searle and Gruber (2017).
- Outside (additional topics, such as generalized linear models, quantile regression, etc.): Select chapters from Khuri (2010), Fahrmeir et al. (2013), Agresti (2015).

3) One-semester course: Univariate Time-Series Analysis.
   - Core: Chapters 4, 5, 6, and 7, and Appendix A.
   - Add: Chapters 8, 9, and 10, and Appendix B.
   - Outside: Select chapters from Brockwell and Davis (2016), Pesaran (2015), Rachev et al. (2007).
4) Two-semester course: Time-Series Analysis.
   - Core: Chapters 4, 5, 6, 7, 8, 9, 10, and 11, and Appendices A and B.
   - Add: Chapters 12 and 13, and Appendix C.
   - Outside (for spectral analysis, VAR, and Kalman filtering): Select chapters from Hamilton (1994), Pollock (1999), Lütkepohl (2005), Tsay (2014), Brockwell and Davis (2016).
   - Outside (for econometric topics such as GMM, use of instruments, and simultaneous equations): Select chapters from Hayashi (2000), Pesaran (2015), Greene (2017).
5) One-semester course: Multivariate Financial Returns Modeling and Portfolio Optimization.
   - Core (not this book): Chapters 5 and 9 (univariate mixed normal, and tail estimation) from book III.
   - Core: Chapters 10, 11, 12, 13, and 14, and Appendix C.
   - Add: Chapter 5 (for TVP regression such as for the CAPM).
   - Outside: Select chapters from Alexander (2008), Jondeau et al. (2007), Rachev et al. (2007), Tsay (2010), Tsay (2012), and Zivot (2018).[3]
6) Mini-course on SAS.
   Appendix D is on data manipulation and basic usage of the SAS system. This is admittedly an oddity, as I use Matlab throughout (as a matrix-based prototyping language) as opposed to a primarily canned-procedure package, such as SAS, SPSS, Minitab, Eviews, Stata, etc.
   The appendix serves as a tutorial on the SAS system, written in a relaxed, informal way, walking the reader through numerous examples of data input, manipulation, and merging, and use of basic statistical analysis procedures. It is included as I believe SAS still has its strengths, as discussed in its opening section, and will be around for a long time. I demonstrate its use for ANOVA in Chapters 2 and 3. As with spoken languages, knowing more than one is often useful, and in this case being fluent in one of the prototyping languages, such as Matlab, R, Python, etc., and one of (if not the arguably most important) canned-routine/data processing languages, is a smart bet for aspiring data analysts and researchers.

In line with books I, II, and III, attention is explicitly paid to application and numeric computation, with examples of Matlab code throughout. The point of including code is to offer a framework for discussion and illustration of numerics, and to show the "mapping" from theory to computation,

2 All these books are excellent in scope and suitability for the numerous topics associated with applied regression analysis, including case studies with real data. It is part of the reason this author sees no good reason to attempt to improve upon them. Notable is Graybill and Iyer (1994) for their emphasis on prediction, and use of confidence intervals (for prediction and model parameters) as opposed to hypothesis tests; see my diatribe in Chapter III.2.8 supporting this view.

3 Jondeau et al. (2007) provides a toolbox of Matlab programs, while Tsay (2012) and Zivot (2018) do so for R.

in contrast to providing black-box programs for an applied user to run when analyzing a data set. Thus, the emphasis is on algorithmic development for implementations involving number crunching with vectors and matrices, as opposed to, say, linking to financial or other databases, string handling, text parsing and processing, generation of advanced graphics, machine learning, design of interfaces, use of object-oriented programming, etc.. As such, the choice of Matlab should not be a substantial hindrance to users of, say, R, Python, or (particularly) Julia, wishing to port the methods to their preferred platforms. A benefit of those latter languages, however, is that they are free. The reader without access to Matlab but wishing to use it could use GNU Octave, which is free, and has essentially the same format and syntax as Matlab.

The preface of book III contains acknowledgements to the handful of professors with whom I had the honor of working, and who were highly instrumental in "forging me" as an academic, as well as to the numerous fellow academics and students who kindly provided me with invaluable comments and corrections on earlier drafts of this book, and book III. Specific to this book, master's student (!!) Christian Frey gets the award for "most picky" (in a good sense), having read various chapters with a very fine-toothed comb, alerting me to numerous typos and unclarities, and also indicating numerous passages where "a typical master's student" might enjoy a bit more verbosity in explanation. Chris also assisted me in writing (the harder parts of) Sections 1.A and C.2. I would give him an honorary doctorate if I could. I am also highly thankful to the excellent Wiley staff who managed this project, as well as copy editor Lesley Montford, who checked every chapter and alerted me to typos, inconsistencies, and other aspects of the presentation, leading to a much better final product. I (grudgingly) take blame for any further errors.

Part I

Linear Models: Regression and ANOVA

1

The Linear Model

The application of econometrics requires more than mastering a collection of tricks. It also requires insight, intuition, and common sense.

(Jan R. Magnus, 2017, p. 31)

The natural starting point for learning about statistical data analysis is with a sample of independent and identically distributed (hereafter i.i.d.) data, say $\mathbf{Y} = (Y_1, \ldots, Y_n)$, as was done in book III. The *linear regression model* relaxes both the identical and independent assumptions by (i) allowing the means of the Y_i to depend, in a linear way, on a set of other variables, (ii) allowing for the Y_i to have different variances, and (iii) allowing for correlation between the Y_i.

The linear regression model is not only of fundamental importance in a large variety of quantitative disciplines, but is also the basis of a large number of more complex models, such as those arising in panel data studies, time-series analysis, and generalized linear models (GLIM), the latter briefly introduced in Section 1.6. Numerous, more advanced data analysis techniques (often referred to now as algorithms) also have their roots in regression, such as the *least absolute shrinkage and selection operator* (LASSO), the *elastic net*, and *least angle regression* (LARS). Such methods are often now showcased under the heading of machine learning.

1.1 Regression, Correlation, and Causality

It is uncomfortably true, although rarely admitted in statistics texts, that many important areas of science are stubbornly impervious to experimental designs based on randomisation of treatments to experimental units. Historically, the response to this embarrassing problem has been to either ignore it or to banish the very notion of causality from the language and to claim that the shadows dancing on the screen are all that exists.

Ignoring the problem doesn't make it go away and defining a problem out of existence doesn't make it so. We need to know what we can safely infer about causes from their observational shadows, what we can't infer, and the degree of ambiguity that remains.

(Bill Shipley, 2016, p. 1)[1]

1 The metaphor to dancing shadows goes back a while, at least to Plato's Republic and the Allegory of the Cave. One can see it today in shadow theater, popular in Southeast Asia; see, e.g., Pigliucci and Kaplan (2006, p. 2).

Linear Models and Time-Series Analysis: Regression, ANOVA, ARMA and GARCH, First Edition. Marc S. Paolella.

The univariate linear regression model relates the scalar random variable Y to k other (possibly random) variables, or **regressors**, $x_1, \dots, x_k$ in a linear fashion,

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon, \tag{1.1}$$

where, typically, $\epsilon \sim \mathrm{N}(0, \sigma^2)$. Values $\beta_1, \dots, \beta_k$ and σ^2 are unknown, constant parameters to be estimated from the data. A more useful notation that also emphasizes that the means of the Y_i are not constant is

$$Y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \epsilon_i, \quad i = 1, 2, \dots, n, \tag{1.2}$$

where now a double subscript on the regressors is necessary. The ϵ_i represent the difference between the values of Y_i and the model used to represent them, $\sum_{j=1}^{k} \beta_j x_{i,j}$, and so are referred to as the **error terms**. It is important to emphasize that the error terms are i.i.d., but the Y_i are not. However, if we take $k = 1$ and $x_{i,1} \equiv 1$, then (1.2) reduces to $Y_i = \beta_1 + \epsilon_i$, which is indeed just the i.i.d. model with $Y_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\beta_1, \sigma^2)$. In fact, it is usually the case that $x_{i,1} \equiv 1$ for any $k \geqslant 1$, in which case the model is said to **include a constant** or **have an intercept term**.

We refer to Y as the **dependent** (random) variable. In other contexts, Y is also called the **endogenous** variable, while the k regressors can also be referred to as the **explanatory**, **exogenous**, or **independent** variables, although the latter term should not be taken to imply that the regressors, when viewed as random variables, are necessarily independent from one another.

The linear structure of (1.1) is one way of building a relationship between the Y_i and a set of variables that "influence" or "explain" them. The usefulness of establishing such a relationship or **conditional** model for the Y_i can be seen in a simple example: Assume a demographer is interested in the income of people living and employed in Hamburg. A random sample of n individuals could be obtained using public records or a phone book, and (rather unrealistically) their incomes Y_i, $i = 1, \dots, n$, elicited. Assuming that income is approximately normally distributed, an **unconditional** model for income could be postulated as $\mathrm{N}(\mu_u, \sigma_u^2)$, where the subscript u denotes the unconditional model and the usual estimators for the mean and variance of a normal sample could be used.

(We emphasize that this example is just an excuse to discuss some concepts. While actual incomes for certain populations can be "reasonably" approximated as Gaussian, they are, of course, not: They are strictly positive, will thus have an extended right tail, and this tail might be heavy, in the sense of being Pareto—this naming being no coincidence, as Vilfredo Pareto worked on modeling incomes, and is also the source of what is now referred to in micro-economics as Pareto optimality. An alternative type of linear model, referred to as GLIM, that uses a non-Gaussian distribution instead of the normal, is briefly discussed below in Section 1.6. Furthermore, interest might not center on modeling the mean income—which is what regression does—but rather the median, or the lower or upper quantiles. This leads to quantile regression, also briefly discussed in Section 1.6.)

A potentially much more precise description of income can be obtained by taking certain factors into consideration that are highly related to income, such as age, level of education, number of years of experience, gender, whether he or she works part or full time, etc. Before continuing this simple example, it is imperative to discuss the three Cs: correlation, causality, and control.

Observe that (simplistically here, for demonstration) age and education might be positively correlated, simply because, as the years go by, people have opportunities to further their schooling and training. As such, if one were to claim that income tends to increase as a function of age, then one cannot conclude this arises out of "seniority" at work, but rather possibly because some of the older people

have received more schooling. Another way of saying this is, while income and age are positively correlated, an increase in age is not necessarily **causal** for income; age and income may be **spuriously correlated**, meaning that their correlation is driven by other factors, such as education, which might indeed be causal for income. Likewise, if one were to claim that income tends to increase with educational levels, then one cannot claim this is due to education *per se*, but rather due simply to seniority at the workplace, possibly despite their enhanced education. Thus, it is important to include both of these variables in the regression.

In the former case, if a positive relationship is found between income and age *with education also in the regression*, then one can conclude a seniority effect. In the literature, one might say "Age appears to be a significant predictor of income, and this being concluded after having also **controlled for** education." Examples of controlling for the relevant factors when assessing causality are ubiquitous in empirical studies of all kinds, and are essential for reliable inference. As one example, in the field of "economics and religion" (which is now a fully established area in economics; see, e.g., McCleary, 2011), in the abstract of one of the highly influential papers in the field, Gruber (2005) states "Religion plays an important role in the lives of many Americans, but there is relatively little study by economists of the implications of religiosity for economic outcomes. This likely reflects the enormous difficulty inherent in separating the causal effects of religiosity from other factors that are correlated with outcomes." The paper is filled with the expression "having controlled for".

A famous example, in a famous paper, is Leamer (1983, Sec. V), showing how conclusions from a study of the factors influencing the murder rate are highly dependent on which set of variables are included in the regression. The notion of controlling for the right variables is often the vehicle for critiquing other studies in an attempt to correct potentially wrong conclusions. For example, Farkas and Vicknair (1996, p. 557) state "[Cancio et al.] claim that discrimination, measured as a residual from an earnings attainment regression, increased after 1976. Their claim depends crucially on which variables are controlled and which variables are omitted from the regression. We believe that the authors have omitted the key control variable—cognitive skill."

The concept of causality is fundamental in econometrics and other social sciences, and we have not even scratched the surface. The different ways it is addressed in popular econometrics textbooks is discussed in Chen and Pearl (2013), and debated in Swamy et al. (2015), Raunig (2017), and Swamy et al. (2017). These serve to indicate that the theoretical framework for understanding causality and its interface to statistical inference is still developing. The importance of causality for scientific inquiry cannot be overstated, and continues to grow in importance in light of artificial intelligence. As a simple example, humans understand that weather is (global warming aside) exogenous, and carrying an umbrella does not cause rain. How should a computer know this? Starting points for further reading include Pearl (2009), Shipley (2016), and the references therein.

Our development of the linear model in this chapter serves two purposes: First, it is the required theoretical statistical framework for understanding ANOVA models, as introduced in Chapters 2 and 3. As ANOVA involves designed experiments and randomization, as opposed to observational studies in the social sciences, we can avoid the delicate issues associated with assessing causality. Second, the linear model serves as the underlying structure of autoregressive time-series models as developed in Part II, and our emphasis is on statistical forecasting, as opposed to the development of structural economic models that explicitly need to address causality.

We now continue with our very simple illustration, just to introduce some terminology. Let $x_{i,2}$ denote the age of the ith person. A conditional model with a constant and age as a regressor is given by $Y_i = \beta_1 + \beta_2 x_{i,2} + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma^2)$. The intercept is measured by β_1 and the slope of income

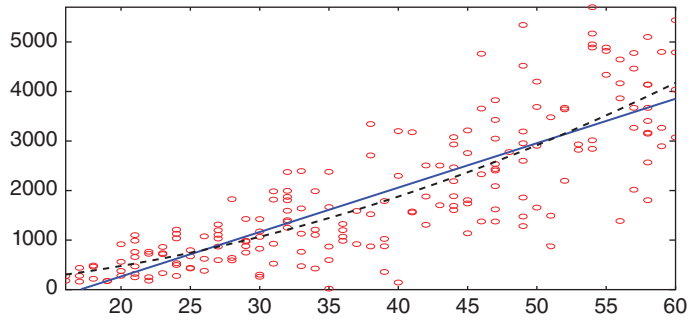


Figure 1.1 Scatterplot of age versus income overlaid with fitted regression curves.

is measured by β_2. Because age is expected to explain a considerable part of variability in income, we expect σ^2 to be significantly less than σ_u^2. A useful way of visualizing the model is with a scatterplot of $x_{i,2}$ and y_i. Figure 1.1 shows such a graph based on a fictitious set of data for 200 individuals between the ages of 16 and 60 and their monthly net income in euros. It is quite clear from the scatterplot that age and income are positively correlated. If age is neglected, then the i.i.d. normal model for income results in $\hat{\mu}_u = 1{,}797$ euros and $\hat{\sigma}_u = 1{,}320$ euros. Using the techniques discussed below, the regression model gives estimates $\hat{\beta}_1 = -1{,}465$, $\hat{\beta}_2 = 85.4$, and $\hat{\sigma} = 755$, the latter being about 43% smaller than $\hat{\sigma}_u$. The model implies that, conditional on the age x, the income Y is modeled as $\mathrm{N}(-1{,}465 + 85.4x, 755^2)$. This is valid only for $16 \leqslant x \leqslant 60$; because of the negative intercept, small values of age would erroneously imply a negative income. The fitted model $y = \hat{\beta}_1 + \hat{\beta}_2 x$ is overlaid in the figure as a solid line.

Notice in Figure 1.1 that the linear approximation underestimates income for both low and high age groups, i.e., income does not seem perfectly linear in age, but rather somewhat quadratic. To accommodate this, we can add another regressor, $x_{i,3} = x_{i,2}^2$, into the model, i.e., $Y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \sigma_q^2)$ and σ_q^2 denotes the conditional variance based on the quadratic model. It is important to realize that the model is still linear (in the constant, age, and age squared). The fitted model turns out to be $Y_i = 190 - 12.5x_{i,2} + 1.29x_{i,3}$, with $\hat{\sigma}_q = 733$, which is about 3% smaller than $\hat{\sigma}$. The fitted curve is shown in Figure 1.1 as a dashed line.

One caveat still remains with the model for income based on age: The variance of income appears to increase with age. This is a typical finding with income data and agrees with economic theory. It implies that both the mean and the variance of income are functions of age. In general, when the variance of the regression error term is not constant, it is said to be **heteroskedastic**, as opposed to **homoskedastic**. The generalized least squares extension of the linear regression model discussed below can be used to address this issue when the structure of the heteroskedasticity as a function of the $\mathbf{X}$ matrix is known.

In certain applications, the ordering of the dependent variable and the regressors is important because they are observed in time, usually equally spaced. Because of this, the notation Y_t will be used, $t = 1, \ldots, T$. Thus, (1.2) becomes

$$Y_t = \beta_1 x_{t,1} + \beta_2 x_{t,2} + \cdots + \beta_k x_{t,k} + \epsilon_t, \quad t = 1, 2, \ldots, T,$$

where $x_{t,i}$ indicates the tth observation of the ith explanatory variable, $i = 1, \ldots, k$, and ϵ_t is the tth error term. In standard matrix notation, the model can be compactly expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.3}$$

where $[\mathbf{X}]_{t,i} = x_{t,i}$, i.e., with $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,k})'$,

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_T' \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & & \vdots \\ x_{T,1} & x_{T,2} & & x_{T,k} \end{bmatrix}, \quad \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}),$$

$\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are $T \times 1$, $\mathbf{X}$ is $T \times k$ and $\boldsymbol{\beta}$ is $k \times 1$. The first column of $\mathbf{X}$ is usually $\mathbf{1}$, the column of ones. Observe that $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$.

An important special case of (1.3) is with $k = 2$ and $x_{t,1} = 1$. Then $Y_t = \beta_1 + \beta_2 X_t + \epsilon_t$, $t = 1, \ldots, T$, is referred to as the **simple linear regression model**. See Problems 1.1 and 1.2.

1.2 Ordinary and Generalized Least Squares

1.2.1 Ordinary Least Squares Estimation

The most popular way of estimating the k parameters in $\boldsymbol{\beta}$ is the **method of least squares**,[2] which takes $\widehat{\boldsymbol{\beta}} = \arg\min \mathsf{S}(\boldsymbol{\beta})$, where

$$\mathsf{S}(\boldsymbol{\beta}) = \mathsf{S}(\boldsymbol{\beta}; \mathbf{Y}, \mathbf{X}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \sum_{t=1}^{T} (Y_t - \mathbf{x}_t'\boldsymbol{\beta})^2, \tag{1.4}$$

and we suppress the dependency of S on $\mathbf{Y}$ and $\mathbf{X}$ when they are clear from the context.

Assume that $\mathbf{X}$ is of full rank k. One procedure to obtain the solution, commonly shown in most books on regression (see, e.g., Seber and Lee, 2003, p. 38), uses matrix calculus; it yields $\partial\, \mathsf{S}(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, and setting this to zero gives the solution

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{1.5}$$

This is referred to as the **ordinary least squares**, or o.l.s., estimator of $\boldsymbol{\beta}$. (The adjective "ordinary" is used to distinguish it from what is called generalized least squares, addressed in Section 1.2.3 below.) Notice that $\widehat{\boldsymbol{\beta}}$ is also the solution to what are referred to as the **normal equations**, given by

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}. \tag{1.6}$$

To verify that (1.5) indeed corresponds to the minimum of $\mathsf{S}(\boldsymbol{\beta})$, the second derivative is checked for positive definiteness, yielding $\partial^2\, \mathsf{S}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}' = 2\mathbf{X}'\mathbf{X}$, which is necessarily positive definite when $\mathbf{X}$ is full rank. Observe that, if $\mathbf{X}$ consists only of a column of ones, which we write as $\mathbf{X} = \mathbf{1}$, then $\widehat{\boldsymbol{\beta}}$ reduces to the mean, $\bar{Y}$, of the Y_t. Also, if $k = T$ (and $\mathbf{X}$ is full rank), then $\widehat{\boldsymbol{\beta}}$ reduces to $\mathbf{X}^{-1}\mathbf{Y}$, with $\mathsf{S}(\widehat{\boldsymbol{\beta}}) = 0$.

Observe that the derivation of $\widehat{\boldsymbol{\beta}}$ in (1.5) did not involve any explicit distributional assumptions. One consequence of this is that the estimator may not have any meaning if the maximally existing moment of the $\{\epsilon_t\}$ is too low. For example, take $\mathbf{X} = \mathbf{1}$ and $\{\epsilon_t\}$ to be i.i.d. Cauchy; then $\widehat{\beta} = \bar{Y}$ is a useless estimator. If we assume that the first moment of the $\{\epsilon_t\}$ exists and is zero, then, writing $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$, we see that $\widehat{\boldsymbol{\beta}}$ is unbiased:

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{\beta}. \tag{1.7}$$

2 This terminology dates back to Adrien-Marie Legendre (1752–1833), though the method is most associated in its origins with Carl Friedrich Gauss, (1777–1855). See Stigler (1981) for further details.

Next, if we have existence of second moments, and $\mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$, then $\mathbb{V}(\widehat{\boldsymbol{\beta}} \mid \sigma^2)$ is given by

$$\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \sigma^2] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}. \tag{1.8}$$

It turns out that $\widehat{\boldsymbol{\beta}}$ has the smallest variance among all linear unbiased estimators; this result is often referred to as the **Gauss–Markov Theorem**, and expressed as saying that $\widehat{\boldsymbol{\beta}}$ is the best linear unbiased estimator, or BLUE. We outline the usual derivation, leaving the straightforward details to the reader. Let $\widehat{\boldsymbol{\beta}}^* = \mathbf{A}'\mathbf{Y}$, where $\mathbf{A}'$ is a $k \times T$ nonstochastic matrix (it can involve $\mathbf{X}$, but not $\mathbf{Y}$). Let $\mathbf{D} = \mathbf{A} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. First calculate $\mathbb{E}[\widehat{\boldsymbol{\beta}}^*]$ and show that the unbiased property implies that $\mathbf{D}'\mathbf{X} = \mathbf{0}$. Next, calculate $\mathbb{V}(\widehat{\boldsymbol{\beta}}^* \mid \sigma^2)$ and show that $\mathbb{V}(\widehat{\boldsymbol{\beta}}^* \mid \sigma^2) = \mathbb{V}(\widehat{\boldsymbol{\beta}} \mid \sigma^2) + \sigma^2\mathbf{D}'\mathbf{D}$. The result follows because $\mathbf{D}'\mathbf{D}$ is obviously positive semi-definite and the variance is minimized when $\mathbf{D} = \mathbf{0}$.

In many situations, it is reasonable to assume normality for the $\{\epsilon_t\}$, in which case we may easily estimate the $k+1$ unknown parameters σ^2 and β_i, $i = 1, \ldots, k$, by maximum likelihood. In particular, with

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi\sigma^2)^{-T/2} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}, \tag{1.9}$$

and log-likelihood

$$\ell(\boldsymbol{\beta}, \sigma^2; \mathbf{Y}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\,\mathsf{S}(\boldsymbol{\beta}), \tag{1.10}$$

where $\mathsf{S}(\boldsymbol{\beta})$ is given in (1.4), setting

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = -\frac{2}{2\sigma^2}\mathbf{X}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \quad \text{and} \quad \frac{\partial \ell}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}\,\mathsf{S}(\boldsymbol{\beta})$$

to zero yields the same estimator for $\boldsymbol{\beta}$ as given in (1.5) and $\tilde{\sigma}^2 = \mathsf{S}(\widehat{\boldsymbol{\beta}})/T$. It will be shown in Section 1.3.2 that the maximum likelihood estimator (hereafter m.l.e.) of σ^2 is biased, while estimator

$$\hat{\sigma}^2 = \mathsf{S}(\widehat{\boldsymbol{\beta}})/(T - k) \tag{1.11}$$

is unbiased.

As $\widehat{\boldsymbol{\beta}}$ is a linear function of $\mathbf{Y}$, $(\widehat{\boldsymbol{\beta}} \mid \sigma^2)$ is multivariate normally distributed, and thus characterized by its first two moments. From (1.7) and (1.8), it follows that $(\widehat{\boldsymbol{\beta}} \mid \sigma^2) \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

1.2.2 Further Aspects of Regression and OLS

> The coefficient of multiple determination, R^2, is a measure many statisticians love to hate. This animosity exists primarily because the widespread use of R^2 inevitably leads to at least occasional misuse.
>
> (Richard Anderson-Sprecher, 1994)

In general, the quantity $\mathsf{S}(\widehat{\boldsymbol{\beta}})$ is referred to as the **residual sum of squares**, abbreviated RSS. The **explained sum of squares**, abbreviated ESS, is defined to be $\sum_{t=1}^{T}(\widehat{Y}_t - \bar{Y})^2$, where the *fitted value* of Y_t is $\widehat{Y}_t := \mathbf{x}_t'\widehat{\boldsymbol{\beta}}$, and the **total (corrected) sum of squares**, or TSS, is $\sum_{t=1}^{T}(Y_t - \bar{Y})^2$. (Annoyingly, both words "error" and "explained" start with an "e", and some presentations define SSE to be the error sum of squares, which is our RSS; see, e.g., Ravishanker and Dey, 2002, p. 101.)

The term *corrected* in the TSS refers to the adjustment of the Y_t for their mean. This is done because the mean is a "trivial" regressor that is not considered to do any real explaining of the dependent variable. Indeed, the total *uncorrected* sum of squares, $\sum_{t=1}^{T} Y_t^2$, could be made arbitrarily large just by adding a large enough constant value to the Y_t, and the model consisting of just the mean (i.e., an $\mathbf{X}$ matrix with just a column of ones) would have the appearance of explaining an arbitrarily large amount of the variation in the data.

While certainly $Y_t - \bar{Y} = (Y_t - \widehat{Y}_t) + (\widehat{Y}_t - \bar{Y})$, it is not immediately obvious that

$$\sum_{t=1}^{T} (Y_t - \bar{Y})^2 = \sum_{t=1}^{T} (Y_t - \widehat{Y}_t)^2 + \sum_{t=1}^{T} (\widehat{Y}_t - \bar{Y})^2,$$

i.e.,

$$\text{TSS} = \text{RSS} + \text{ESS}. \tag{1.12}$$

This fundamental identity is proven below in Section 1.3.2.

A popular statistic that measures the fraction of the variability of $\mathbf{Y}$ taken into account by a linear regression model that includes a constant, compared to use of just a constant (i.e., $\bar{Y}$), is the **coefficient of multiple determination**, designated as R^2, and defined as

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\mathsf{S}(\widehat{\boldsymbol{\beta}}, \mathbf{Y}, \mathbf{X})}{\mathsf{S}(\bar{Y}, \mathbf{Y}, \mathbf{1})}, \tag{1.13}$$

where $\mathbf{1}$ is a T-length column of ones. The coefficient of multiple determination R^2 provides a measure of the extent to which the regressors "explain" the dependent variable over and above the contribution from just the constant term. It is important that $\mathbf{X}$ contain a constant or a set of variables whose linear combination yields a constant; see Becker and Kennedy (1992) and Anderson-Sprecher (1994) and the references therein for more detail on this point.

By construction, the observed R^2 is a number between zero and one. As with other quantities associated with regression (such as the nearly always reported "t-statistics" for assessing individual "significance" of the regressors), R^2 is a statistic (a function of the data but not of the unknown parameters) and thus *is a random variable*. In Section 1.4.4 we derive the F test for parameter restrictions. With J such linear restrictions, and $\widehat{\boldsymbol{\gamma}}$ referring to the restricted estimator, we will show (1.88), repeated here, as

$$F = \frac{[\mathsf{S}(\widehat{\boldsymbol{\gamma}}) - \mathsf{S}(\widehat{\boldsymbol{\beta}})]/J}{\mathsf{S}(\widehat{\boldsymbol{\beta}})/(T-k)} \sim \text{F}(J, T-k), \tag{1.14}$$

under the null hypothesis H_0 that the J restrictions are true. Let $J = k - 1$ and $\widehat{\boldsymbol{\gamma}} = \bar{Y}$, so that the restricted model is that all regressor coefficients, *except the constant* are zero. Then, comparing (1.13) and (1.14),

$$F = \frac{T-k}{k-1} \frac{R^2}{1-R^2}, \qquad \text{or} \qquad R^2 = \frac{(k-1)F}{(T-k)+(k-1)F}. \tag{1.15}$$

Dividing the numerator and denominator of the latter expression by $T - k$ and recalling the relationship between F and beta random variables (see, e.g., Problem I.7.20), we immediately have that

$$R^2 \sim \text{Beta}\left(\frac{k-1}{2}, \frac{T-k}{2}\right), \tag{1.16}$$

so that $\mathbb{E}[R^2] = (k-1)/(T-1)$ from, for example, (I.7.12). Its variance could similarly be stated. Recall that its distribution was derived under the null hypothesis that the $k-1$ regression coefficients are zero. This implies that R^2 is upward biased, and also shows that just adding superfluous regressors will always increase the expected value of R^2. As such, choosing a set of regressors such that R^2 is maximized is not appropriate for model selection.

However, the so-called **adjusted** R^2 can be used. It is defined as

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{T-1}{T-k}. \tag{1.17}$$

Virtually all statistical software for regression will include this measure. Less well known is that it has (like so many things) its origin with Ronald Fisher; see Fisher (1925). Notice how, like the Akaike information criterion (hereafter AIC) and other penalty-based measures applied to the obtained log likelihood, when k is increased, the increase in R^2 is offset by a factor involving k in R^2_{adj}.

Measure (1.17) can be motivated in (at least) two ways. First, note that, under the null hypothesis,

$$\mathbb{E}[R^2_{\text{adj}}] = 1 - \left(1 - \frac{k-1}{T-1}\right)\frac{T-1}{T-k} = 0,$$

providing a perfect offset to R^2's expected value simply increasing in k under the null. A second way is to note that, while $R^2 = 1 - \text{RSS}/\text{TSS}$ from (1.13),

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(T-k)}{\text{TSS}/(T-1)} = 1 - \frac{\widehat{\mathbb{V}}(\widehat{\boldsymbol{\epsilon}})}{\widehat{\mathbb{V}}(\mathbf{Y})},$$

the numerator and denominator being unbiased estimators of their respective variances, recalling (1.11). The use of R^2_{adj} for model selection is very similar to use of other measures, such as the (corrected) AIC and the so-called **Mallows'** C_k; see, e.g., Seber and Lee (2003, Ch. 12) for a very good discussion of these, and other criteria, and the relationships among them.

Section 1.2.3 extends the model to the case in which $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ from (1.3), but $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a known, positive definite variance–covariance matrix. There, an appropriate expression for R^2 will be derived that generalizes (1.13). For now, the reader is encouraged to express R^2 in (1.13) as a ratio of quadratic forms, assuming $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, and compute and plot its density for a given $\mathbf{X}$ and $\boldsymbol{\Sigma}$, such as given in (1.31) for a given value of parameter a, as done in, e.g., Carrodus and Giles (1992). When $a = 0$, the density should coincide with that given by (1.16).

We end this section with an important remark, and an important example.

Remark It is often assumed that the elements of $\mathbf{X}$ are known constants. This is quite plausible in designed experiments, where $\mathbf{X}$ is chosen in such a way as to maximize the ability of the experiment to answer the questions of interest. In this case, $\mathbf{X}$ is often referred to as the **design matrix**. This will rarely hold in applications in the social sciences, where the $\mathbf{x}_t'$ reflect certain measurements and are better described as being observations of random variables from the multivariate distribution describing both $\mathbf{x}_t'$ and Y_t. Fortunately, under certain assumptions, one may ignore this issue and proceed as if $\mathbf{x}_t'$ were fixed constants and not realizations of a random variable.

Assume matrix $\mathbf{X}$ is no longer deterministic. Denote by $\mathbf{X}$ an outcome of random variable $\mathcal{X}$, with kT-variate probability density function (hereafter p.d.f.) $f_{\mathcal{X}}(\mathbf{X}\,;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a parameter vector. We require the following assumption:

0. The conditional distribution $\mathbf{Y} \mid (\mathcal{X} = \mathbf{X})$ depends only on $\mathbf{X}$ and parameters $\boldsymbol{\beta}$ and σ and such that $\mathbf{Y} \mid (\mathcal{X} = \mathbf{X})$ has mean $\mathbf{X}\boldsymbol{\beta}$ and finite variance $\sigma^2\mathbf{I}$.

For example, we could have $\mathbf{Y} \mid (\mathcal{X} = \mathbf{X}) \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Under the stated assumption, the joint density of $\mathbf{Y}$ and $\mathcal{X}$ can be written as

$$f_{\mathbf{Y},\mathcal{X}}(\mathbf{y}, \mathbf{X} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = f_{\mathbf{Y}|\mathcal{X}}(\mathbf{y} \mid \mathbf{X}; \boldsymbol{\beta}, \sigma^2) \cdot f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}). \tag{1.18}$$

Now consider the following two additional assumptions:

1) The distribution of $\mathcal{X}$ does not depend on $\boldsymbol{\beta}$ or σ^2, so we can write $f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\theta})$.
2) The parameter space of $\boldsymbol{\theta}$ and that of $(\boldsymbol{\beta}, \sigma^2)$ are not related, that is, they are not restricted by one another in any way.

Then, with regard to $\boldsymbol{\beta}$ and σ^2, $f_{\mathcal{X}}$ is only a multiplicative constant and the log-likelihood corresponding to (1.18) is the same as (1.10) plus the additional term $\log f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\theta})$. As this term does not involve $\boldsymbol{\beta}$ or σ^2, the (generalized) least squares estimator still coincides with the m.l.e. When the above assumptions are satisfied, $\boldsymbol{\theta}$ and $(\boldsymbol{\beta}, \sigma^2)$ are said to be **functionally independent** (Graybill, 1976, p. 380), or **variation-free** (Poirier, 1995, p. 461). More common in the econometrics literature is to say that one assumes $\mathbf{X}$ to be **(weakly) exogenous** with respect to $\mathbf{Y}$.

The extent to which these assumptions are reasonable is open to debate. Clearly, without them, estimation of $\boldsymbol{\beta}$ and σ^2 is not so straightforward, as then $f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ must be (fully, or at least partially) specified. If they hold, then

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \mathbb{E}_{\mathcal{X}}[\mathbb{E}[\widehat{\boldsymbol{\beta}} \mid \mathcal{X} = \mathbf{X}]] = \mathbb{E}_{\mathcal{X}}[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}[\boldsymbol{\epsilon} \mid \mathcal{X}]] = \mathbb{E}_{\mathcal{X}}[\boldsymbol{\beta}] = \boldsymbol{\beta}$$

and

$$\mathbb{V}(\widehat{\boldsymbol{\beta}} \mid \sigma^2) = \mathbb{E}_{\mathcal{X}}[\mathbb{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \mathcal{X} = \mathbf{X}, \sigma^2]] = \sigma^2\mathbb{E}_{\mathcal{X}}[(\mathcal{X}'\mathcal{X})^{-1}],$$

the latter being obtainable only when $f_{\mathcal{X}}(\mathbf{X}; \boldsymbol{\theta})$ is known.

A discussion of the implications of falsely assuming that $\mathbf{X}$ is not stochastic is provided by Binkley and Abbott (1987).[3] ■

Example 1.1 ***Frisch–Waugh–Lovell Theorem***

It is occasionally useful to express the o.l.s. estimator of each component of the partitioned vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$, where $\boldsymbol{\beta}_1$ is $k_1 \times 1, 1 \leqslant k_1 < k$. With the appropriate corresponding partition of $\mathbf{X}$, model (1.3) is then expressed as

$$\mathbf{Y} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}.$$

The normal equations (1.6) then read

$$\begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \end{pmatrix} \mathbf{Y},$$

or

$$\mathbf{X}_1'\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{Y} \qquad \text{and} \qquad \mathbf{X}_2'\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1 + \mathbf{X}_2'\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{Y}, \tag{1.19}$$

3 We use the tombstone, QED, or halmos, symbol ■ to denote the end of proofs of theorems, as well as examples and remarks, acknowledging that it is traditionally only used for the former, as popularized by Paul Halmos.

so that

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{Y} - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2) \tag{1.20}$$

and $\widehat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'(\mathbf{Y} - \mathbf{X}_1\widehat{\boldsymbol{\beta}}_1)$. To obtain an expression for $\widehat{\boldsymbol{\beta}}_2$ that does not depend on $\widehat{\boldsymbol{\beta}}_1$, let $\mathbf{M}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$, premultiply (1.20) by $\mathbf{X}_1$, and substitute $\mathbf{X}_1\widehat{\boldsymbol{\beta}}_1$ into the second equation in (1.19) to get

$$\mathbf{X}_2'(\mathbf{I} - \mathbf{M}_1)(\mathbf{Y} - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2) + \mathbf{X}_2'\mathbf{X}_2\widehat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{Y},$$

or, expanding and solving for $\widehat{\boldsymbol{\beta}}_2$,

$$\widehat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y}. \tag{1.21}$$

A similar argument (or via symmetry) shows that

$$\widehat{\boldsymbol{\beta}}_1 = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y}, \tag{1.22}$$

where $\mathbf{M}_2 = \mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'$.

An important special case of (1.21) discussed further in Chapter 4 is when $k_1 = k - 1$, so that $\mathbf{X}_2$ is $T \times 1$ and $\widehat{\boldsymbol{\beta}}_2$ in (1.21) reduces to the scalar

$$\widehat{\beta}_2 = \frac{\mathbf{X}_2'\mathbf{M}_1\mathbf{Y}}{\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2}. \tag{1.23}$$

This is a ratio of a bilinear form to a quadratic form, as discussed in Appendix A.

The Frisch–Waugh–Lovell theorem has both computational value (see, e.g., Ruud, 2000, p. 66, and Example 1.9 below) and theoretical value; see Ruud (2000), Davidson and MacKinnon (2004), and also Section 5.2. Extensions of the theorem are considered in Fiebig et al. (1996). ■

1.2.3 Generalized Least Squares

Now consider the more general assumption that $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a known, positive definite variance–covariance matrix. The density of $\mathbf{Y}$ is now given by

$$f_{\mathbf{Y}}(\mathbf{y}) = (2\pi)^{-T/2}|\sigma^2\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}, \tag{1.24}$$

and one could use calculus to find the m.l.e. of $\boldsymbol{\beta}$. Alternatively, we could transform the model in such a way that the above results still apply. In particular, with $\boldsymbol{\Sigma}^{-1/2}$ the symmetric matrix such that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$, premultiply (1.3) by $\boldsymbol{\Sigma}^{-1/2}$ so that

$$\boldsymbol{\Sigma}^{-1/2}\mathbf{Y} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon}, \qquad \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon} \sim \mathrm{N}_T(\mathbf{0}, \sigma^2\mathbf{I}). \tag{1.25}$$

Then, using the previous maximum likelihood approach as in (1.10), with

$$\mathbf{Y}_* := \boldsymbol{\Sigma}^{-1/2}\mathbf{Y} \quad \text{and} \quad \mathbf{X}_* := \boldsymbol{\Sigma}^{-1/2}\mathbf{X} \tag{1.26}$$

in place of $\mathbf{Y}$ and $\mathbf{X}$ implies the normal equations

$$(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = \mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \tag{1.27}$$

that generalize (1.6), and

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Sigma}} = (\mathbf{X}_*'\mathbf{X}_*)^{-1}\mathbf{X}_*'\mathbf{Y}_* = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}, \tag{1.28}$$