Praise for earlier printings

"A beautiful introduction to the theory of econometrics, because of its careful selection of topics, its lucid writing, and its good humor. It is a perfect textbook for students who already know some matrix algebra and statistics and who wish to learn the essentials of econometric theory."

Jaap Abbring, Tilburg University

"Econometrics is a difficult subject to learn and to teach. This book makes both aspects fun and easy. A must-have book to use and learn from."

Jesús Gonzalo, Universidad Carlos III de Madrid

"A concise and entertaining introduction into the essentials of econometric theory. The text is very accessible to well-trained undergraduates and I highly recommend it to anyone interested in a more thorough treatment of the subject than what is typically covered in undergraduate courses."

Guido Kuersteiner, University of Maryland

"This book is student friendly and designed to meet the needs of conveying the key fundamental results in econometrics. A textbook of this kind is long overdue. It breaks the tradition and prevailing wisdom of keeping matrix algebra from undergraduate students. Magnus's textbook may be the beginning of a new era in teaching and learning econometrics."

Rachida Ouysse, University of New South Wales

"This is a carefully written text which is ideal for an honors course in econometrics in U.S. undergraduate economics programs. It has many entertaining examples and its non-intimidating length suits a single-semester course."

Joris Pinkse, Pennsylvania State University

"Beautifully written by a true master of his craft. Although terse, the discussion is always insightful and uncluttered by extraneous detail. [. . .] Highly recommended."

Christopher L. Skeels, The University of Melbourne
Economic Record, vol. 94

"A compact and approachable matrix-based introduction to econometric theory based around the classical normal linear model. The coverage is extensive and wide-ranging in topics dealing with least squares and maximum likelihood estimation, inference and prediction with discussions of exact and large sample theory and conditioning. The appendices contain useful and compact discussions of standard matrix analysis and statistical results."

Richard Smith, University of Cambridge

"This is a fascinatingly short and excellent introduction to classical econometrics, teaching the key ideas and insights and the key theorems and derivations, without getting side-tracked, all illuminated with delightful little cartoons and stories. Start here, and then venture to the tomb of your choice for further details! Highly recommended."

Harald Uhlig, University of Chicago

"A remarkable book. It provides the central ideas and methods of econometrics using a uniquely fresh, concise, and interesting writing style. It is not written for a general audience, but if you come to an econometrics course with a basic knowledge of matrix algebra and statistics, then this is the book for you."

Xinyu Zhang, Chinese Academy of Sciences

Introduction to the theory of econometrics

Jan R. Magnus

Emeritus Professor, Tilburg University
Extraordinary Professor, Vrije Universiteit Amsterdam

VU University Press, Amsterdam

VU University Press
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands

www.vuuniversitypress.com
info@vuuitgeverij.nl

First printing, April 2017
Second printing (with minor corrections), August 2017
Third printing (with minor corrections), February 2018

Type setting (in LaTeX): Jan R. Magnus
Cartoons: Joan Berkhemer
Cover design: Marion Rosendahl

ISBN 978 90 8659 766 6
NUR 789

To the next generation:
Dorian, Midas, Tybalt, and Julian

Interview with the author

by E. D. Young

YOUNG: You have written a book.

MAGNUS: I wrote several.

YOUNG: I mean your recent little introductory book. I read it.

MAGNUS: Did you like it?

YOUNG: No.

MAGNUS: I am sorry to hear that. Do you want to know why I wrote the book?

YOUNG: Yes, it puzzles me.

MAGNUS: After my retirement from Tilburg University in the Summer of 2013, I accepted a position as Extraordinary Professor at the Vrije Universiteit in Amsterdam. Here I was asked to teach a course in econometric theory for second-year undergraduate students. I had not taught such a course for many years, so this was a challenge.

YOUNG: I can imagine. Especially at your age.

MAGNUS: In their first year these students learn some calculus, matrix algebra, and statistics, but no econometrics. Their first encounter with econometrics takes place in the first term of their second year. This twelve-week course consists of a two-hour lecture and a two-hour exercise class per week. The first six weeks of the course don't use matrix algebra; the second half does. I was asked to give the lectures (not the exercises) for the second half.

YOUNG: So you looked for a suitable textbook.

MAGNUS: Precisely. But I did not find one. When I was a student in the late 1960s and early 1970s there were many textbooks providing econometric theory for students with a proper mathematical and statistical background. Now, there are few.

YOUNG: One is enough.

MAGNUS: True, but I find the few modern textbooks often too big, too overwhelming, and too dull. Although econometrics is now a large field, I believe that the collection of central ideas and methods is actually quite small. The aim in my book is to concentrate on those ideas and methods, and to show that econometrics is exciting.

YOUNG: Perhaps you are right that there are not many books now at the more advanced undergraduate level. But there are many books providing econometric theory without matrix algebra.

MAGNUS: Yes, and some of these are very good (most are not). But my students did come to the course with a knowledge of matrix algebra and statistics, and this knowledge is useful in understanding econometrics. So I wanted to provide a course that catered for their level of knowledge.

YOUNG: So the little book contains the lectures you gave for this group of students?

MAGNUS: More or less.

YOUNG: Apart from the silly stories you tell, the book seems to reflect your personal interests and preoccupations.

MAGNUS: I suppose it does. But it also contains much of the standard material that you would expect at an introductory level. I believe that two things are essential in such a course, because they provide the basis for understanding important generalizations such as instrumental variables, generalized method of moments, and extremum estimators: first, a thorough knowledge of the standard linear regression model; and second, a thorough understanding of the principles of maximum likelihood. The chapters in the book reflect this belief.

YOUNG: I have compiled a list of all the things that you do *not* talk about: autocorrelation and heteroskedasticity, Bayesian econometrics, random regressors, time se...

MAGNUS: You are right, the book is not comprehensive. In fact, it is quite the opposite. The challenge in writing it was what to delete, not what to add. I believe it is better to teach less material if at least, at the end of the course, the student can not only reproduce some of the material but is able to actually work with it and apply the theory to solve his or her personal econometric problem. This is only possible if the student learns the basic ideas underlying econometric thought.

YOUNG: In a way the book is rather old-fashioned. It could have been written fifty years ago.

MAGNUS: It is interesting you say that. The basics of econometrics have not changed that much. Most of the material that I

discuss is also in Theil's *Principles of Econometrics*, which came out in 1971 and served as a masterly guide when I and my fellow students studied econometrics.

YOUNG: Does this mean we learned nothing new in the past fifty years?

MAGNUS: Of course not. Econometrics has developed rapidly in many directions, but the basics are still more or less intact.

YOUNG: Has the field only grown or have there also been subjects that have become irrelevant?

MAGNUS: The prime example of a subject that was important in the past is simultaneous equations. This is hardly taught any more. Other things were unknown in the past and are popular now, for example cointegration, unit roots, and model averaging. These new directions are not suitable for an introductory course and should be discussed in later courses, for final-year undergraduates or graduate students.

YOUNG: I notice that you do discuss model averaging in this little book.

MAGNUS: Yes, it is one of my hobbies. But I also included it because model averaging brings out an aspect that is becoming increasingly important in econometrics, namely the distinction between focus and auxiliary variables.

YOUNG: This is hardly new. Ed Leamer already wrote about it in the late 1970s.

MAGNUS: Leamer's ideas did not have much effect on how we teach econometrics. Traditionally the main task of applied econometrics was to explain economic variables, and this was achieved by estimating equations with many explanatory variables, all supposedly of equal importance. But these days empirical research is more interested in estimating causal effects, hence focusing on one or two specific regressors (the focus regressors). The other regressors (the auxiliary regressors) are less important; their role is to improve the estimation of the focus parameters, but the associated auxiliary parameter values are not of primary interest. This distinction between focus and auxiliary variables and parameters plays a key role in my book.

YOUNG: How about truth? You don't seem to have a high regard for truth.

MAGNUS: In an econometric context, the truth represents the process that generated the observations, some would say 'the true model'. In the old days this is what we were told to discover. Once we knew the truth, we could use it in many directions: estimate

parameters, forecast future values, or recommend policies. We now know that this is not feasible. A model is an approximation of the truth and a different approximation is required depending on the question you wish to answer. This implies that we should more often ask 'does it matter?' rather than 'is it true?' The current book emphasizes this distinction at various points.

YOUNG: You added two appendices to the book. Why?

MAGNUS: Not all students have precisely the same background and some will have forgotten some of the material that they should know. The two appendices contain brief reminders of what I require as background knowledge in matrix algebra and statistics. Of course, I had to draw the line somewhere. So, for example, I assume that the student is familiar with the univariate normal distribution, but not necessarily with the multivariate normal distribution. And I assume that the student can differentiate, but not that he or she is familiar with vector or matrix calculus.

YOUNG: What is your hope for this book?

MAGNUS: Simply that it will be of use to a future generation of economists and econometricians and will generate some joy and enthusiasm.

YOUNG: Finally, do you have any recommendations for this future generation?

MAGNUS: Perhaps one, namely to compete for the Philip Swallow Award of which I am currently the sole recipient.

YOUNG: Philip Swallow? Is he not a character in David Lodge's *Changing Places*? I did not know there was an award named after him.

MAGNUS: There is. There is so much emphasis now on counting published pages in prestigious journals. The more pages you have and the more prestigious the journal, the more points you get and this plays a role in tenure decisions and promotions. I do not like this. In the end what matters is the content, not the number of pages or the name of the journal. In order to voice my concern I instigated an annual award for the person whose paper had been rejected by the lowest-ranked journal. In the end only one person ever applied for this award, namely me.

YOUNG: No doubt your most treasured honor.

Amsterdam/Wapserveen
March/August 2017

Acknowledgements

I am grateful to my students and to Rutger Lit, who taught the exercise class accompanying the course, for their many comments. My colleagues Chico Blasques, Peter Boswijk, Lennart Hoogerheide, Franc Klaassen, Siem Jan Koopman, Bertrand Melenberg, Franco Peracchi, Anatoly Peresetsky, Götz Trenkler, and Andrey Vasnev provided valuable and constructive feedback on earlier versions. And Eveline de Jong invented some of the stories that E. D. Young finds so silly.

I benefitted from the book review in *Economic Record* (2018) by Chris Skeels. This review is rather positive, but it also contains two points of critique. In preparing the third printing, I have taken the opportunity to give Chapter 2 a different focus (in particular Sections 2.8–2.10) in order to accommodate the first of Chris' well-justified criticisms, but I was unable to accommodate his second criticism regarding Chapter 4 without introducing more statistical machinery than I think is appropriate for a text at this level.

Some of the stories are taken from my own previously published papers. Figure 1.2 on page 3 is taken from Ikefuji *et al.* (2015). The story about Ottolenghi's dish on page 32 is inspired by a similar story (concerning music rather than food) which appeared in De Luca *et al.* (2017). The remote island story on page 35 comes from Magnus (1999). I thank the journals for letting me use this material again.

The figures were first drawn by Charles Bos, Rutger Lit, and Giuseppe de Luca; then redrawn and 'professionalized' by Charles Bos. The violinist Joan Berkhemer produced the cartoons and Piet van Oostrum patiently answered all my LaTeX questions. My publisher, Jan Oegema, made the transition from text to book a joyful and stimulating experience. I am grateful to all for their positive and inspiring input.

Contents

1 | Approximation

1.1 The captain's age

Imagine a ship, but not an ordinary ship. This is an ocean liner powered by four diesel engines and two additional gas turbines. It has a length of 345 meters, weighs 79,300 tons, and has 14,000 square meters of exterior deck space. There are eighteen decks, four swimming pools, and many restaurants. The ship's ocean speed is 30 knots (56 km/h; 35 mph) and it houses 2500 passengers and 1250 officers and crew. Now, what is the age of the captain?

At first glance, there seems little we can say about the captain's age. But, upon reflection, maybe we can. To be captain of such a flagship one needs much experience, so the captain cannot be too young. One also needs to be fit and healthy, so the captain cannot be too old either. It is fairly safe to estimate the captain's age therefore as in-between 45 and 55. In fact, the average age of the captains of the Queen Mary 2 was 47 over the last ten years. (This was the story I and my fellow students were told on our first lecture in econometrics. It was supposedly an answer to the question: what is econometrics? And it is not such a bad story either. In econometrics we constantly struggle with the fact that the model is never quite right and the data are never quite what we want.)

1.2 A straight line

Econometrics is often thought of as a branch of statistics, analyzing uncertain events. In this first chapter, however, nothing is uncertain. I simply ask the following question: given a collection of points, how do we draw the best line through these points. This is a question of approximation rather than of estimation. We shall be concerned with estimation in the next chapter.

A straight line is usually written as $y = a + bx$, where a and b are called the *parameters* of the line. Instead of a and b, I shall write β_1 and β_2, because later I shall need many more βs. Thus, the equation

$$y = \beta_1 + \beta_2 x$$

describes a linear relationship (that is, a line) relating y to x. If we know the two parameters, then we can calculate y for any x we like.

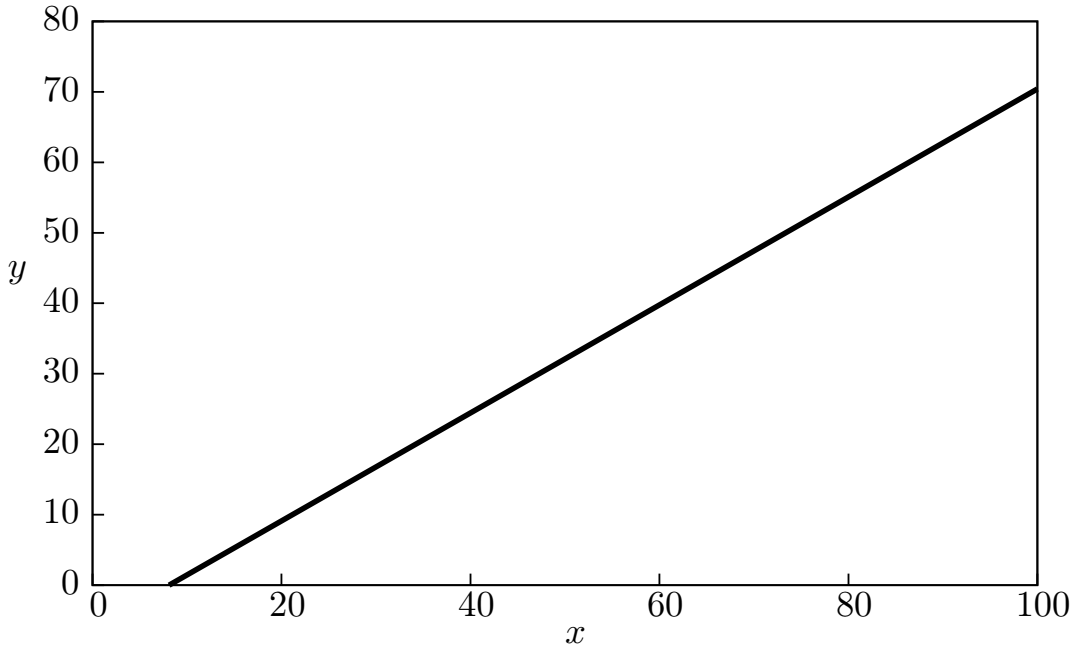


Figure 1.1: A straight line

A straight line, like the one in Figure 1.1, is the simplest example of a model. Suppose we believe that the higher your income the more you will spend on food. That is not an unreasonable assumption. But what does the relationship between income and food consumption look like? Is it a line? And, if so, is it precisely a line or only approximately?

In Figure 1.2 I added some points. I collected observations on $n = 100$ families. For the ith family I know their income x_i and the amount of money y_i they spend on food. This gives me n points (x_i, y_i). I don't expect these n points to lie *exactly* on a line, but if my model (a straight line) is to be a good model then the points should lie *approximately* on the line.

This seems to be the case here. For incomes in the center, roughly between 40 and 80, the consumption function can be well

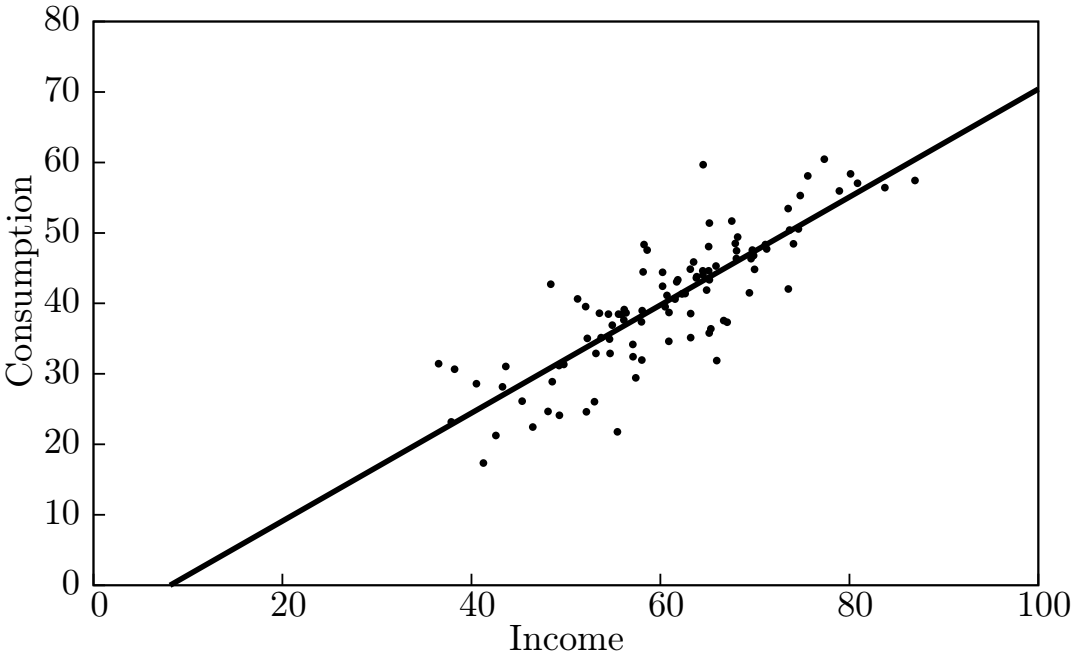


Figure 1.2: A consumption function

approximated by the line. To make further progress we need to be more precise. Thus we ask the following two questions:

- How do we measure 'closeness' to the line?
- What do we mean when we say that an approximation is 'good'? In other words, how close to the line should the points be in order that a straight line is a good model?

Both points will be discussed in this chapter, but before doing so I point out that we may well ask a third question, namely: what is the practical significance of this line for very low and very high incomes? In other words, does the line describe the consumption-income relationship for *all* incomes or just for those in the center? The answer is clear. For very poor and very wealthy families the linear relationship fails to work well. For very low incomes, predicted consumption would even become negative! This does not mean that a linear consumption function is useless, but it is only useful in the center of the domain. Models are approximations, not truths, and approximations may not work well if we move too far away from the point of approximation. This is so in all sciences. In physics, for example, Newton's theory works fine for cars and trains, but not for space ships.

1.3 Least squares

Let's consider the first question first: how do we measure closeness? If we have n points (x_i, y_i) and these points would lie precisely on a line, then y_i could be written as $y_i = \beta_1 + \beta_2 x_i$. In reality, the points will not lie exactly on the line and there will be deviations, so I write

$$y_i = \beta_1 + \beta_2 x_i + u_i \qquad (i = 1, \ldots, n), \tag{1.1}$$

where the u_i denote deviations from the line. If a deviation u_i is zero, then the corresponding point lies precisely on the line, but in general most or all deviations will not be zero.

I can write (1.1) in matrix form as follows:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

or, for short,

$$y = X\beta + u. \tag{1.2}$$

Our problem is to choose β such that the deviations u_i are as 'small' as possible. There are several ways to define 'small' in this context. The most common (and most important) is to consider the function

$$\phi(\beta) = \sum_{i=1}^{n} u_i^2 = u'u = (y - X\beta)'(y - X\beta),$$

and minimize this function with respect to β. This method is called *least squares* (LS). Let us write

$$\begin{aligned} \phi(\beta) &= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta \\ &= y'y - 2y'X\beta + \beta'X'X\beta, \end{aligned}$$

where we use the fact that $y'X\beta = \beta'X'y$ because the transpose of a scalar (1×1 matrix) is the same scalar. Then, by the differentiation rules in Section A.11 of Appendix A,

$$\frac{\partial \phi(\beta)}{\partial \beta'} = -2y'X + 2\beta'X'X.$$

Setting this derivative equal to zero and transposing, we find the solution b by solving

$$X'Xb = X'y,$$

and this gives

$$b = (X'X)^{-1}X'y, \tag{1.3}$$

assuming that $X'X$ is nonsingular (invertible) or, what is the same, that X has full column rank; see Section A.3.

The solution (1.3) is the least-squares approximation. Given the data y and X, we may now construct a new variable, say

$$\hat{y} = Xb = X(X'X)^{-1}X'y, \tag{1.4}$$

which we can call the *predictor* of y and which, by construction, lies precisely on the line.

In the above development I assumed that the X matrix has only two columns (simple regression), but in finding the least-squares approximation I did not use this fact at all. In general, the X matrix will have k columns corresponding to k regressors (multiple regression). The equation $y = X\beta + u$, written out in full, would then look like this:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

The element x_{ij} denotes the ith observation on the jth regressor. The first regressor is usually a vector of ones (the constant term), but this is not necessary.

1.4 An alternative derivation

Whatever the expression for b, I can always decompose $u = y - X\beta$ as

$$y - X\beta = (y - Xb) + X(b - \beta).$$

Such a decomposition is particularly useful when the two component vectors are orthogonal to each other. (Two vectors a and b are said to be orthogonal when $a'b = 0$.) This is the case here, because

$$X'(y - Xb) = X'y - X'X(X'X)^{-1}X'y = X'y - X'y = 0. \tag{1.5}$$

Hence,

$$(y - X\beta)'(y - X\beta) = (y - Xb)'(y - Xb) + (b - \beta)'X'X(b - \beta),$$

which is clearly minimized by choosing $\beta = b$.

This proof is simpler because it does not need differentiation, but it is less appealing, because it is not constructive: you need to know the answer before you start the proof.

1.5 Residuals

Where $u = y - X\beta$ denotes the vector of deviations, I now define the vector of *residuals* as

$$e = y - Xb = y - X(X'X)^{-1}X'y = My, \tag{1.6}$$

where

$$M = I_n - X(X'X)^{-1}X' \tag{1.7}$$

is a symmetric idempotent matrix. (A square matrix A is idempotent if it satisfies $AA = A$. Idempotent matrices need not be symmetric, although in practice most are; see Appendix A.7.) The deviation vector u is a function of the variable β. If we choose β 'optimally' so that $\beta = b$, then we obtain the residuals e, which thus depend only on X and y. Since

$$\begin{aligned} MX &= (I_n - X(X'X)^{-1}X')X \\ &= X - X(X'X)^{-1}X'X = X - X = 0, \end{aligned} \tag{1.8}$$

we find

$$X'e = X'My = 0, \tag{1.9}$$

confirming what we have already seen in (1.5).

In most cases, the X matrix contains a vector of ones (the constant term). If this is the case, then the sum of the residuals is zero. This follows, because the fact that $X'e = 0$ shows that *each* of the k regressors is orthogonal to the vector of residuals e. In particular, if there is a constant term, then one of the regressors (typically the first) is $\imath$, the vector of ones, and hence

$$\sum_{i=1}^{n} e_i = \imath'e = 0. \tag{1.10}$$

1.6 Geometry

There is a different way to obtain the least-squares solution, not by algebra but by geometry. Consider two points p and q on a line. The distance between p and q is defined as

$$d = |p - q| = \sqrt{(p - q)^2}.$$

The distance between two points $p = (p_1, p_2)$ and $q = (q_1, q_2)$ on a plane is defined as

$$d = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} = \sqrt{(p - q)'(p - q)},$$

and, similarly, the distance between p and q in the n-dimensional Euclidean space $\mathbb{R}^n$ is given by

$$d = \sqrt{(p - q)'(p - q)}.$$

Let $S(X)$ be the k-dimensional subspace of $\mathbb{R}^n$ containing all vectors that are linear combinations of the columns of X, that is, can be represented as $X\beta$ for some β in $\mathbb{R}^k$. Now consider two points in $\mathbb{R}^n$, namely y and a linear combination $X\beta$ of the columns of X. The distance between these two points is

$$\sqrt{(y - X\beta)'(y - X\beta)},$$

and the shortest distance is obtained when we choose β equal to $b = (X'X)^{-1}X'y$ so that Xb is orthogonal to $y - Xb$.

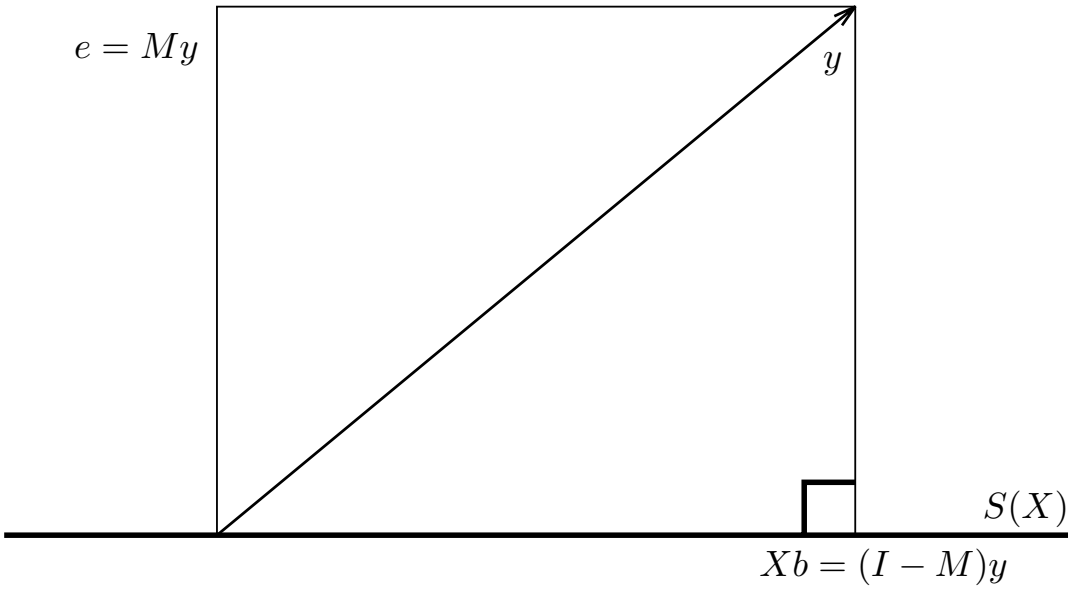


Figure 1.3: Geometric interpretation of least squares

In Figure 1.3 we see this demonstrated in two dimensions. The k-dimensional plane $S(X)$ is represented by the horizontal line. The vector y is projected onto the space $S(X)$ of the regressors to obtain the linear combination Xb that is as close as possible to y, and we see that y decomposes as $y = Xb + e$, where $Xb = (I - M)y$ and $e = y - Xb = My$ are orthogonal to each other.

1.7 Least absolute deviations

Least squares were introduced in Section 1.3 as a means to define 'closeness' to a line. This is not the only possible definition of closeness. Instead of minimizing the sum of the squared deviations we could also minimize the sum of the absolute deviations,

$$\phi(\beta) = \sum_{i=1}^{n} |u_i|.$$

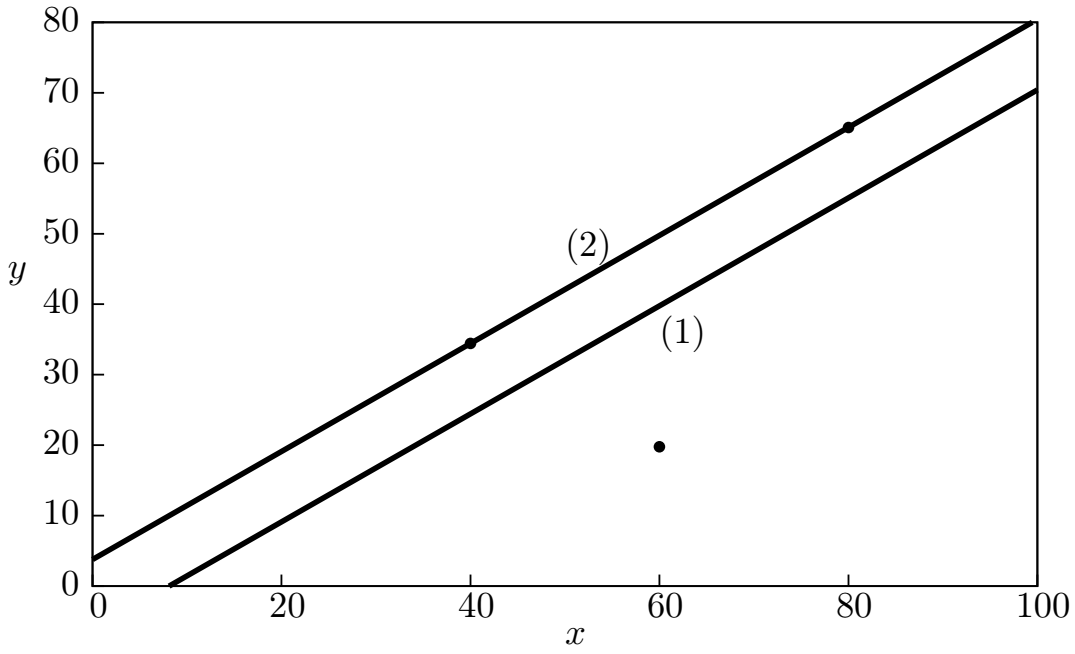


Figure 1.4: Least absolute deviations with three points

The traditional argument against least absolute deviations is illustrated in Figure 1.4. Here we have only three points at $x_1 = 40$, $x_2 = 60$, and $x_3 = 80$. The line labeled (1) is the least-squares line, the same as in the previous figures. The points y_1 and y_3 lie 10 points above the line, while y_2 lies 20 points below the line, so that

$$e = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \end{pmatrix} = \begin{pmatrix} 10 \\ -20 \\ 10 \end{pmatrix}$$

and

$$X'e = \begin{pmatrix} 1 & 1 & 1 \\ 40 & 60 & 80 \end{pmatrix} \begin{pmatrix} 10 \\ -20 \\ 10 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

in accordance with (1.9) and (1.10).

The sum of the absolute deviations is $10 + 20 + 10 = 40$. If we move the line 10 points up, we get line (2). Here the sum of the absolute deviations is $0 + 30 + 0 = 30$, which is 10 points lower then for the line (1). In fact, line (2) is the best possible from the least absolute deviations point of view: there exists no other line where the sum of the three absolute deviations is smaller than 30. (If we consider more than three points then the solution may not be unique.)

Hence, from the point of view of absolute deviations, line (2) is better than line (1). Most people find this counterintuitive, because one would expect the best line to lie somewhere in-between the points and not at the edge. This is one (but not the only) reason why least squares is generally preferred above least absolute deviations. Another reason is that absolute deviations are not differentiable, a huge disadvantage in theoretical derivations.

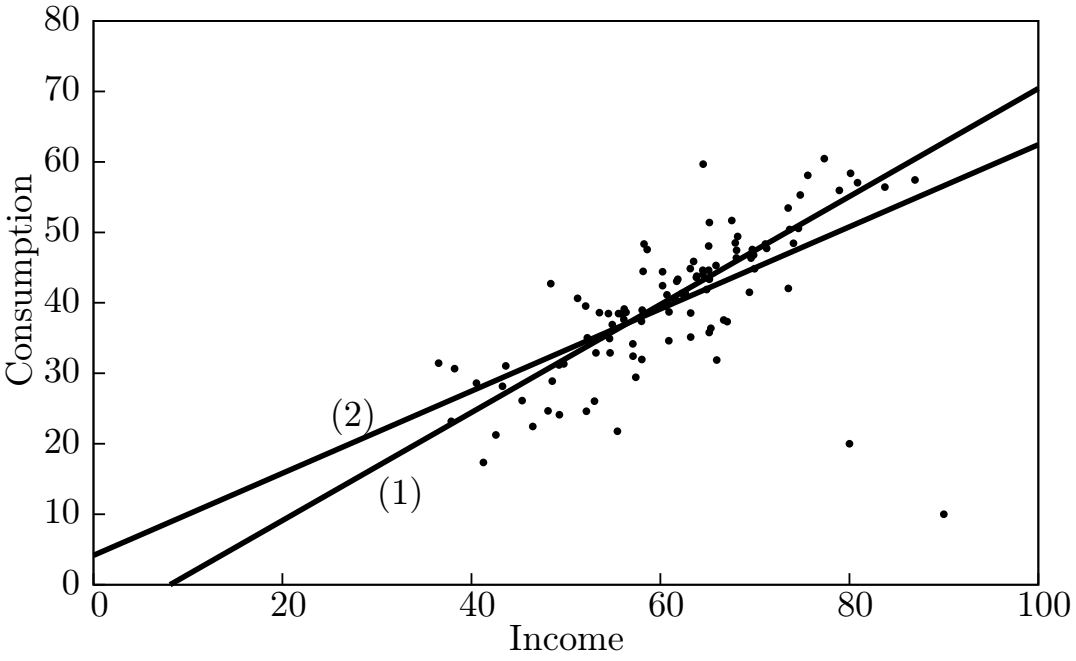


Figure 1.5: Least squares with an outlier

But now consider Figure 1.5. This is the same as Figure 1.2, except that I have added two points, namely $(80, 20)$ and $(90, 10)$. Line (1) is the line without the additional points (hence the same as in Figure 1.2), while line (2) is the line with the additional points. Clearly, there is a difference between the two regression lines, and most statisticians would consider this difference large. Adding one or two points may thus lead to a notably different line. This is also

unsatisfactory.

In summary, least squares provides a unique solution, but it is not robust, while least absolute deviations provides a more robust result, but there are certain disadvantages (nondifferentiability, possibly multiple solutions). Least absolute deviations is robust in that it is resistant to outliers in the data. It gives equal emphasis to all observations, in contrast to least squares which, by squaring the residuals, gives more weight to large residuals. Least squares is still the dominant method, but least absolute deviations has also become popular in the last decade.

1.8 The fit

In Section 1.2 two questions were raised. First, how do we measure 'closeness' to the line? This question was answered in Sections 1.3 and 1.7. Second, what do we mean when we say an approximation is 'good'? Let me discuss this question now.

By the definition of residuals we have

$$y = Xb + e = \hat{y} + e,$$

where $\hat{y} = Xb$ is the predictor. Since $X'e = 0$, we have $\hat{y}'e = b'X'e = 0$ and hence

$$y'y = \hat{y}'\hat{y} + e'e.$$

Therefore, the ratio

$$\frac{\hat{y}'\hat{y}}{y'y} = 1 - \frac{e'e}{y'y}$$

lies between zero and one, and it provides the fraction of the variation of the dependent variable y that can be attributed to the variation in the explanatory variables X. If the ratio is close to one, then this signifies that the fitted value is close to the dependent variable.

In most cases there is a constant term, so that the residuals add up to zero. Let me introduce the matrix

$$A = I_n - \frac{\imath\imath'}{n},$$

where $\imath$ denotes a vector of ones. This is a symmetric idempotent matrix, because

$$AA = \left(I_n - \frac{\imath\imath'}{n}\right)\left(I_n - \frac{\imath\imath'}{n}\right) = I_n - \frac{\imath\imath'}{n} - \frac{\imath\imath'}{n} + \frac{\imath\imath'}{n} = A,$$

since $\imath'\imath = n$. (In fact, A is a special case of the matrix M defined in (1.7) by taking $X = \imath$.) Now, if there is a constant term, then $\imath'e = 0$ and $Ae = e$, so that

$$Ay = AXb + Ae = AXb + e,$$

and hence, since $(AX)'e = X'A'e = X'Ae = X'e = 0$,

$$y'Ay = b'X'AXb + e'e. \tag{1.11}$$

We express this equality as

$$\text{SST} = \text{SSE} + \text{SSR}, \tag{1.12}$$

that is, the total sum of squares (SST) equals the sum of the explained sum of squares (SSE) and the sum of squared residuals (SSR). We emphasize that equality only holds when there is a constant term.

The fit of the least-squares approximation is typically assessed by the *coefficient of multiple determination*,

$$R^2 = 1 - \frac{\text{SSR}}{\text{SST}} = 1 - \frac{e'e}{y'Ay}. \tag{1.13}$$

Provided there is a constant term among the regressors, we have $0 \leq R^2 \leq 1$, but if there is no constant term then R^2 can become negative.

The coefficient R^2 is sensitive to the magnitudes of n and k in small samples, and an adjusted R^2 will be introduced later, in Section 3.14.

1.9 Adding and omitting variables

We have chosen k regressors to approximate y. What would happen if we choose $k - 1$ or, more generally, $k_1 < k$ regressors. It is easy to see that the fit would not be as good.

Let us write $y = X\beta + u$ as

$$y = X_1\beta_1 + X_2\beta_2 + u, \tag{1.14}$$

where $X = (X_1 : X_2)$ is partitioned into an $n \times k_1$ matrix X_1 and an $n \times k_2$ matrix X_2 with $k = k_1 + k_2$. Similarly, the $k \times 1$ vector β is partitioned into two subvectors β_1 and β_2.

If we only use X_1 to approximate y, then we would minimize

$$\phi_1(\beta_1) = u_1'u_1 = (y - X_1\beta_1)'(y - X_1\beta_1)$$

with respect to β_1. This is the same as to minimize

$$\phi(\beta) = u'u = (y - X\beta)'(y - X\beta)$$

with respect to β under the restriction $\beta_2 = 0$. The minimum of $\phi(\beta)$ under a restriction can never be lower than the minimum without the restriction, and hence the fit obtained using the complete X matrix must be at least as good as the fit obtained using only X_1.

In fact, we can be more precise. If e denotes the vector of residuals from the fit using the complete X matrix, then

$$e'e = (My)'(My) = y'M'My = y'MMy = y'My,$$

because M is symmetric and idempotent ($M' = M = M^2$). Similarly, if e_1 denotes the vector of residuals from the fit using only X_1, then $e_1'e_1 = y'M_1y$, where $M_1 = I_n - X_1(X_1'X_1)^{-1}X_1'$. Hence, using the decomposition (A.5) in Section A.10 of M_1 into two positive semidefinite matrices,

$$M_1 = M + M_1X_2(X_2'M_1X_2)^{-1}X_2'M_1, \tag{1.15}$$

we find

$$e_1'e_1 - e'e = y'M_1y - y'My = e_1'X_2(X_2'M_1X_2)^{-1}X_2'e_1,$$

so that

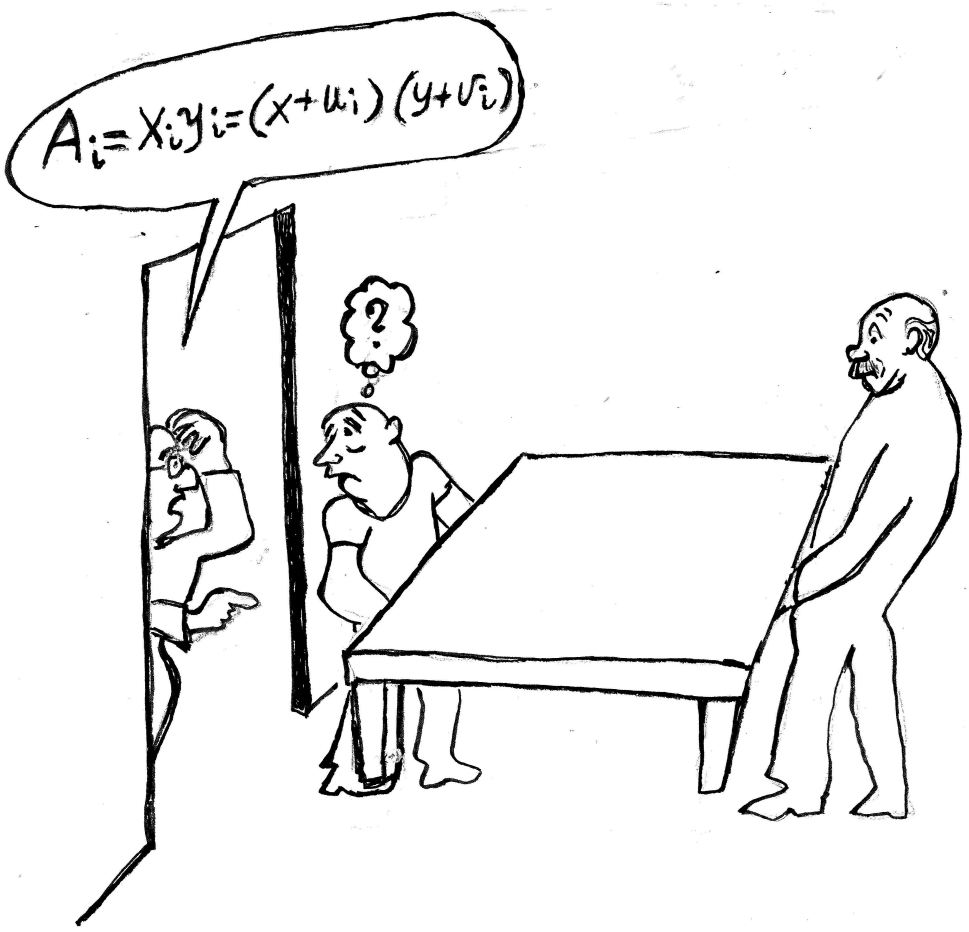
$$e'e \leq e_1'e_1 \qquad \text{with equality if and only if } X_2'e_1 = 0. \tag{1.16}$$

Apparently, there is no point in adding the new regressors X_2 if these are orthogonal to the residuals e_1 obtained from fitting only X_1. The reason is as follows. Let $b = (X'X)^{-1}X'y$ be the solution to the least-squares problem when we fit the complete X matrix. This vector has subvectors b_1 ($k_1 \times 1$) and b_2 ($k_2 \times 1$). We have

$$\begin{aligned} X_2'e_1 = 0 &\iff X_2'M_1y = 0 \\ &\iff b_2 = (X_2'M_1X_2)^{-1}X_2'M_1y = 0, \end{aligned}$$

using (A.4), and hence equality in (1.16) occurs if and only if $b_2 = 0$, that is, if adding X_2 has no effect.

The question whether to add or omit regressors to a model is a key issue in econometric modelling, and I shall return to it at the end of Chapters 2 and 3.


$A_i = X_i Y_i = (x+u_i)(y+v_i)$
?

2 | Best unbiased estimation

2.1 Rothenberg's table

In my office I have a rectangular table. I want to know the length and width of this table, so I hire an assistant to take one hundred measurements (x_i, y_i) of the length x and width y. The assistant believes (wrongly) that I am only interested in the area A of the table, so she multiplies each pair of measurements to obtain one hundred measurements of the area, $A_i = x_i y_i$. Then she destroys the underlying measurements of the length and the width, so that the only thing we have is one hundred measurements of the area. Can we still estimate the length and width of the table?

This is indeed possible. Rothenberg (2005) defines measurement errors

$$u_i = x_i - x, \qquad v_i = y_i - y,$$

and then expresses A_i in terms of these measurement errors,

$$A_i = x_i y_i = (x + u_i)(y + v_i) = xy + xv_i + yu_i + u_i v_i,$$

and similarly for A_i^2 and A_i^3. Then he makes certain (credible) assumptions on these measurement errors and is able to estimate x and y.

There is a compelling byproduct to this story. If the assistant is very sloppy, then of course we would not expect to get good estimates. But if she is perfect (no measurement errors), then this is also bad. In that case we would end up with one hundred identical measurements of the area, so that in effect we only have one observation. There is no way we can estimate the length and width from one observation. Apparently, there exists an optimal level of sloppiness for an assistant and it can be calculated.

2.2 Observations

The word 'observation' is ambiguous. Suppose we do a simple experiment and toss a fair coin one hundred times. Let x denote the fraction of times that heads comes up. Since the coin is fair we would expect x to be close to 0.5, but how close? In this case, we can show that the probability that $0.40 < x < 0.60$ is about 95% and that the probability that $0.37 < x < 0.63$ is about 99%.

This was a theoretical experiment. But now let's actually do it with a real coin. We toss one hundred times and we find 53 heads and 47 tails, so that $x = 0.53$. Now it makes no longer sense to talk about the probability that 0.53 lies between 0.40 and 0.60. It either lies in this interval or it doesn't.

This example demonstrates the difference between a theoretical observation and a realized observation. A theoretical observation is a random variable with a probability distribution, while its realization is a number. Some statistical concepts have different words to distinguish the theoretical concept from its realization. For example, an *estimator* is a random variable, while an *estimate* is its realization. But this is an exception. Most statistical concepts — and 'observation' is one of them — have only one word to denote both. In the previous chapter, observations were realized (given) data, but now they denote theoretical (random) data.

2.3 The linear model

In Chapter 1 the world was nonstochastic and our only task was to approximate a cloud of points by a straight line. (The words 'random' and 'stochastic' are both in common use; they mean the same.) The real world, however, involves many uncertainties. Some of these can be modeled and understood, while some remain completely unpredictable. The simplest model involving stochasticity is the *linear model*, which lies at the heart of econometrics and will be treated in some detail. Of course, the linear model must often be extended in many directions in order to be applicable in a practical situation. The power of the linear model lies herein that it allows such extensions while keeping its key features.

The linear model looks exactly like the approximation equation (1.2),

$$y = X\beta + u,$$

but its interpretation is quite different. In an approximation context, our problem was to choose β such that the deviations u_i are

as small as possible, but in a stochastic framework our problem is to *estimate* β.

The random variable y is called the *dependent* variable on which we have n observations $y_1, \ldots, y_n$, collected in the $n \times 1$ vector y. The dependent variable is related to a set of k *explanatory* variables (or *regressors*) and on each regressor we also have n observations, which leads to an $n \times k$ matrix X. I shall assume, unless otherwise indicated, that the regressors are nonrandom. The $k \times 1$ vector β is a vector of k unknown parameters, and u is a random $n \times 1$ vector whose components are now called *disturbances* (sometimes *errors*).

Note that some variables are random (y and u), while others are nonrandom (X and β), and that some variables are observable (y and X), while others are not observable (β and u).

2.4 Ideal conditions

A model can be seen as a set of restrictions, and the ideal conditions consist of five restrictions. The first restriction is linearity.

Assumption 2.1. *The model is linear.*

What this means is that the model is linear in the β parameters, not that it is linear in X. For example, the equation

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + u_i$$

is a linear model, even though it is not linear in x. If we think of approximating a function $y = f(X)$, then the linear function $y = X\beta$ provides a first-order approximation, so that even if f is in fact not linear, the linear model may still work well in practice.

Assumption 2.2. *The $n \times k$ matrix X is nonrandom and has rank k.*

This assumption repeats that X is assumed to be nonrandom and has full column rank k, which implies that the matrix $X'X$ is nonsingular. (Recall from Section A.3 that X and $X'X$ have the same rank for any matrix X. Hence, if $r(X) = k$, then $r(X'X) = k$ and since $X'X$ is a $k \times k$ matrix, it is nonsingular.)

Assumption 2.3. *The $n \times 1$ vector u has mean zero and variance $\sigma^2 I_n$.*

This means that all components u_i have mean zero, are uncorrelated with each other (that is, $\text{cov}(u_i, u_j) = 0$ for all $i \neq j$), and have

a common variance which we denote by σ^2. We could equivalently have written this assumption as

$$\mathrm{E}(y) = X\beta, \qquad \mathrm{var}(y) = \sigma^2 I_n.$$

Sometimes the formulation in terms of u is more convenient, sometimes the formulation in terms of y.

Assumption 2.4. *The $n \times 1$ vector u follows a multivariate normal distribution.*

Together with the previous assumption, normality implies that the disturbances u_i are not only uncorrelated but in fact independent; see Section B.3. Why should the disturbances follow a normal distribution? I offer three reasons. First, the normal distribution (just as linearity) is easy to work with. This is an important practical consideration, but of course in itself not sufficient as a reason.

Second, the disturbances may be due to errors in measuring y, in which case the normality assumption is reasonable because measurement errors are known to be well approximated by a normal distribution. In fact, the normal distribution was developed in the context of measurement error.

Third, the normal distribution is very much like the assumption of linearity in that it serves as an approximation, in this case a second-order approximation. This can be seen from the appropriate Taylor expansion, as follows. Let $f(u)$ be an arbitrary density function, not necessarily symmetric, with a maximum at $u = 0$ (the mode). Since the derivative of f vanishes at the mode, we obtain from (A.7):

$$\log f(u) \approx \log f(0) + (1/2)u'H(0)u,$$

where the symbol $\approx$ means 'is approximately equal to' and

$$H(u) = \frac{\partial^2 \log f(u)}{\partial u \, \partial u'}$$

is the Hessian matrix. Hence,

$$f(u) \approx f(0) \exp[(-1/2)u'(-H(0))u],$$

which we recognize as the normal distribution with mean zero and variance $[-H(0)]^{-1}$; see (B.2). This tells us that the normal distribution serves as a good approximation to an arbitrary density function, certainly in the center of the distribution.

Assumption 2.5. *If we let* $Q_n = X'X/n$, *then* $Q_n \to Q$ *as* $n \to \infty$, *where* Q *is a finite and positive definite* $k \times k$ *matrix.*

This assumption concerns the asymptotic behavior of the matrix $X'X$. If x_{ij} denotes the ith observation on the jth regressor, then the jth diagonal element of $X'X$ is given by

$$(X'X)_{jj} = \sum_{i=1}^{n} x_{ij}^2.$$

If the first regressor is the constant term, as is often the case, then $(X'X)_{11} = n$ and $(Q_n)_{11} = 1$ for all n. Clearly, the diagonal elements of $X'X$ will increase with n, but they should not increase too fast (in which case some of the elements of Q_n go to ∞) or too slow (in which case Q_n will become singular). We shall not be concerned with asymptotics until Chapter 5.

Normality and linearity thus play similar roles in our setup: linearity is a first-order approximation to the model and normality is a second-order approximation to the distribution.

2.5 Model and data-generation process

Data are generated by some process and we call this process the *data-generation process* (DGP). A model tries to approximate the DGP but it will not be equal to the DGP, unless in truly exceptional and simplified situations.

In what follows I shall usually assume that model and DGP coincide, but not always. For example, the DGP may take the form

$$y = X_1\beta_1 + X_2\beta_2 + u$$

while the model may be

$$y = X_1\beta_1 + u_1.$$

This is a common situation, where reality (the DGP) is more complex than the model we wish or are able to consider.

Perhaps we know that X_2 plays a role but we have no data on these regressors. Then we simply cannot include X_2. But even if we did have data on X_2 we may wish to exclude these regressors from our model. This is somewhat puzzling and counterintuitive. Why would we willingly misspecify the model? I will come back to this important issue in Sections 2.16 and 2.18.

2.6 The usefulness of sloppy notation

It is neither possible nor desirable to be completely accurate in notation. Always one has to compromise between readability and accurateness. This is usually no problem, but it is important to be aware of it.

For example, when I write β it can mean a variable over which we wish to optimize, as in minimizing $(y - X\beta)'(y - X\beta)$. But it can also mean the specific 'true' value of the parameter, which we denote by β_0 if we wish to emphasize it.

This is particularly important when we take expectations. When I write $\mathrm{E}(y) = X\beta$, I always take the expectation with respect to the DGP (the 'true' model), and I should, strictly speaking, write $\mathrm{E}(y) = X\beta_0$, because β here denotes a specific value, corresponding to the DGP.

2.7 Why is X nonrandom?

I have chosen for the assumption that X is nonrandom. This used to be the standard assumption in basic econometrics, but now it is often considered old-fashioned. Hence some defense is called for.

There are situations (for example in experiments) where the researcher can choose the design matrix X, which is then nonrandom by construction. But in most cases it is more realistic and more general to assume that the observations (y, X) are generated by some joint distribution. The assumptions of the previous section must then be adjusted. For example, instead of Assumption 2.3 we would now write

$$\mathrm{E}(u|X) = 0, \qquad \mathrm{var}(u|X) = \sigma^2 I_n.$$

The condition $\mathrm{E}(u|X) = 0$ is sometimes called *(strong) exogeneity*.

Against the advantage of greater generality there is one important disadvantage, namely that we now have to work with conditional distributions and conditional moments. Conditioning is the most difficult part of probability theory. One could argue that without a proper understanding of conditioning one cannot be a good econometrician (and there is much in favor to be said for such a viewpoint), but the basics of econometric theory do not change if we make this simplifying assumption and the material is easier to grasp. This is my defense. I shall discuss conditioning in Chapter 6.

2.8 The principle of best unbiased estimation

Estimation is something quite different than approximation. We wish to construct estimators (in our case, of β and σ^2) that have certain desired sampling properties, given our assumptions. In order to do this we need some guidance on what we mean by a 'good' estimator. In this little book two such guiding principles will be discussed: best unbiased estimation (current chapter) and maximum likelihood (Chapter 4).

The principle of best unbiased estimation is simple and intuitive. Given a parameter θ we consider the class of all unbiased estimators $\hat{\theta}$ of θ, that is, estimators for which $\mathrm{E}(\hat{\theta}) = \theta$, and in this class we choose the one with the smallest variance (that is, the 'best').

In general, the class of unbiased estimators does not have enough structure to find a unique best estimator, and we need to restrict this class further, for example by requiring the distribution to be normal or the estimator to be linear. We shall see examples of this in the sequel.

Unbiasedness is often regarded as a desirable property for an estimator to have, and the current chapter takes unbiasedness as its starting point. For linear estimators unbiasedness make much sense, but for nonlinear estimators this is less obvious. For example, if s^2 is an unbiased estimator of σ^2, what can we say about s as an estimator of σ? This estimator is biased, which follows from *Jensen's inequality*: $f(\mathrm{E}(x)) \leq \mathrm{E}(f(x))$ for any convex function f. Letting $f(x) = x^2$ (convex), we have $(\mathrm{E}\, s)^2 \leq \mathrm{E}(s^2) = \sigma^2$ and hence $\mathrm{E}(s) \leq \sigma$.

If θ is a vector then $\mathrm{var}(\hat{\theta})$ will be a matrix, and we need to define what we mean by saying that one matrix is smaller than another. Thus, for two symmetric matrices A and B, we shall say that A is smaller than (or equal to) B and we write $A \leq B$ (or $B \geq A$) when $B - A$ is positive semidefinite, and we write $A < B$ (or $B > A$) when $B - A$ is positive definite.

2.9 Estimation of a linear combination of the βs

Let's assume that we have a model which coincides with the DGP, however unlikely that may seem, and that Assumptions 2.1–2.3 hold. That is, we consider a model $y = X\beta + u$, where X is nonrandom and of full column-rank and where $\mathrm{E}(u) = 0$ and $\mathrm{var}(u) = \sigma^2 I_n$. We first concentrate on the estimation of β. We shall discuss the estimation of σ^2 in Section 2.12.

The fact that β is a vector rather than a scalar causes some difficulties that will be clarified in the next section. So let's begin by trying to estimate a simple scalar function of the βs, namely a linear combination, say $w'\beta$. The vector w could be anything, for example $w' = (0, 1, 0, 0, \ldots, 0)$ in which case we estimate one component β_2, or $w' = (0, 1, 1, 0, \ldots, 0)$ in which case we estimate the sum $\beta_2 + \beta_3$.

Let $\theta = w'\beta$ be the parameter to be estimated. Since β is a mean parameter (a linear concept), it makes sense to base estimation of θ on a linear function of the data. Thus, we write our estimator as $\hat{\theta} = a'y$, where a is to be chosen 'optimally' in some sense. We are now dealing with a linear estimator in a linear model, and hence it also makes sense to require that this estimator is unbiased (also a linear concept):

$$\mathrm{E}(\hat{\theta}) = \mathrm{E}(a'y) = a'X\beta = w'\beta$$

for all β, implying that $X'a = w$.

This places a restriction on the vector a, but it does not yet fully determine a. We need more, so let's calculate the variance:

$$\mathrm{var}(\hat{\theta}) = \mathrm{var}(a'y) = a'\,\mathrm{var}(y)\,a = a'(\sigma^2 I_n)\,a = \sigma^2 a'a.$$

Suppose now that we can choose a such that the variance of $\hat{\theta}$ is minimized under the restrictions of linearity and unbiasedness. Then we obtain an estimator which is 'best' (minimum variance) in the class of linear unbiased estimators, that is, a best linear unbiased estimator (BLUE). The obtain the BLUE we thus need to minimize $a'a$ subject to the restriction $X'a = w$. This requires Lagrangian theory.

We define the Lagrangian function

$$\psi(a) = a'a/2 - l'(X'a - w)$$

where l is a vector of Lagrangian multipliers. (Note that I write $a'a/2$ rather than $a'a$. This makes no difference, since any a which minimizes $a'a$ will also minimize $a'a/2$, but it is a common trick since we know that we minimize a quadratic function so that a 2 will appear in the derivative. The 1/2 neutralizes this 2.)

The derivative is

$$\frac{\partial \psi}{\partial a'} = a' - l'X',$$

and hence the first-order conditions are

$$a = Xl, \qquad X'a = w.$$

This gives

$$w = X'a = X'Xl,$$

so that $l = (X'X)^{-1}w$ and $a = Xl = X(X'X)^{-1}w$.

We should verify that the solution a indeed minimizes $\phi(a) = a'a/2$ under the constraint $g(a) = X'a - w$. This is easy here, because the constraint g is linear and the function ϕ is strictly convex. This implies that ψ is also strictly convex so that, using the results in Section A.13, $\phi(a)$ attains a strict absolute minimum at the solution $a = X(X'X)^{-1}w$ under the constraint $X'a = w$. Hence, $\hat{\theta} = a'y = w'(X'X)^{-1}X'y$ is BLUE, that is, it has the lowest variance in the class of linear unbiased estimators of θ. Summarizing, we have proved the following result.

Proposition 2.1. *Under Assumptions 2.1–2.3 the estimator $w'\hat{\beta}$ with $\hat{\beta} = (X'X)^{-1}X'y$ is the best linear unbiased estimator (BLUE) of $w'\beta$.*

What we have done is apply the principle of best unbiased estimation to the parameter $w'\beta$ by adding the restriction that the estimator is linear (in y). The fact that $w'\hat{\beta}$ is the BLUE of $w'\beta$ for any choice of the vector w suggests that $\hat{\beta}$ is perhaps the BLUE of β, and this is indeed the case. But it is something that needs to be shown, and we shall do so in the next section.

We recognize $\hat{\beta}$ as the least-squares approximator b of the previous chapter. So, while we started this section by stating that estimation is something quite different than approximation, it turns out that the approximator b of the previous chapter and the estimator $\hat{\beta}$ are in fact identical. This is neither obvious nor trivial. The estimator $\hat{\beta}$ is typically called the 'least-squares' (LS) estimator of β, which is fine as long as we remember that least squares is an approximation principle, not an estimation principle.

2.10 Estimation of β

Now let's try to repeat this analysis for a parameter vector $\theta = W'\beta$, where W is now a matrix rather than a vector. Again, we wish to estimate θ as a linear function of y, say $\hat{\theta} = A'y$, where the matrix A should be chosen 'optimally' in some sense. Imposing unbiasedness gives

$$\mathrm{E}(\hat{\theta}) = \mathrm{E}(A'y) = A'X\beta = W'\beta$$

for all β, implying that $X'A = W$. Using (B.1) the variance is

$$\mathrm{var}(\hat{\theta}) = \mathrm{var}(A'y) = A'\,\mathrm{var}(y)\,A = A'(\sigma^2 I_n)\,A = \sigma^2 A'A.$$

As in the previous section we ask whether we can choose the matrix A such that the variance of $\hat{\theta}$ is minimized under the restrictions of linearity and unbiasedness. If we can, then we have obtained an estimator which is 'best' (minimum variance) in the class of linear unbiased estimators, that is, a best linear unbiased estimator (BLUE).

Before we can answer this question two difficulties need to be resolved. First, what do we mean by saying that a matrix is 'small'? For real scalars we know what we mean by $a \leq b$, but what do we mean by $A \leq B$? This issue was resolved in Section 2.8: a symmetric matrix A is smaller than (or equal to) a symmetric matrix B if and only if $B - A$ is positive semidefinite, and we write this as $A \leq B$ or $B \geq A$.

The second difficulty is that we cannot use calculus to minimize matrix functions. Calculus and Lagrangian theory are designed for scalar functions, not for matrix functions.

But we can still minimize, even without using calculus. We need to minimize $A'A$ subject to the restriction $X'A = W$. Let us define $D = A - X(X'X)^{-1}W$. Then $X'A = W$ if and only if $X'D = 0$, and under this restriction we have

$$\begin{aligned} A'A &= \left(D' + W'(X'X)^{-1}X'\right)\left(D + X(X'X)^{-1}W\right) \\ &= D'D + W'(X'X)^{-1}W \geq W'(X'X)^{-1}W \end{aligned}$$

with equality if and only if $D = 0$. Hence, under the restriction $X'A = W$, the matrix $A'A$ achieves a lower bound and this lower bound is reached when $D = 0$, that is, when $A = X(X'X)^{-1}W$ so that $\hat{\theta} = A'y = W'(X'X)^{-1}X'y = W'\hat{\beta}$. This demonstrates the following extension of Proposition 2.1.

Proposition 2.2. *Under Assumptions 2.1–2.3 the estimator $W'\hat{\beta}$ is the best linear unbiased estimator (BLUE) of $W'\beta$.*

Since this holds for any matrix W, it holds in particular when we choose $W = I_k$.

Proposition 2.3 (Gauss-Markov)**.** *Under Assumptions 2.1–2.3 the least-squares estimator $\hat{\beta} = (X'X)^{-1}X'y$ is the best linear unbiased estimator (BLUE) of β.*

This is the famous Gauss-Markov theorem, named after the German mathematician Johann Carl Friedrich Gauss (1777–1855) and the Russian probabilist Andrey Andreyevich Markov (1856–1922).

It tells us that in the class of linear unbiased estimators there is no estimator better (that is, with lower variance) than $\hat{\beta}$.

Notice that we have restricted the class of unbiased estimators by imposing linearity of the estimator, but we did not need the normality of the disturbances (Assumption 2.4). We shall see later (Section 4.7) that if we assume normality of the disturbances we need not impose linearity of the estimator.

Obviously, $\mathrm{E}(\hat{\beta}) = \beta$, because it was constructed to be unbiased. And indeed, we have

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \mathrm{E}\left((X'X)^{-1}X'y\right) = (X'X)^{-1}X'\,\mathrm{E}(y) \\ &= (X'X)^{-1}X'X\beta = \beta. \end{aligned} \tag{2.1}$$

The variance is

$$\begin{aligned}\mathrm{var}(\hat{\beta}) &= \mathrm{var}\left((X'X)^{-1}X'y\right) = (X'X)^{-1}X'\,\mathrm{var}(y)X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} = \sigma^2(X'X)^{-1}, \end{aligned} \tag{2.2}$$

and we see that the variance is 'large' if there is large variation in the disturbances (large σ^2) or small variation in the regressors (so that $X'X$ is 'small' and $(X'X)^{-1}$ is 'large').

2.11 Residuals again

We encountered residuals in Section 1.5, but we reconsider them here, because now they are random. In accordance with (1.6) we have

$$e = y - X\hat{\beta} = My = Mu, \tag{2.3}$$

where we recall from Section 1.5 that $M = I_n - X(X'X)^{-1}X'$ and $MX = 0$, so that $X'e = 0$. This gives

$$\mathrm{E}(e) = \mathrm{E}(Mu) = 0 \tag{2.4}$$

and

$$\mathrm{var}(e) = \mathrm{var}(Mu) = M\,\mathrm{var}(u)M = \sigma^2 M. \tag{2.5}$$

The residuals are constrained by $X'e = 0$ and hence their variance matrix cannot have full rank. This is confirmed by (2.5) since M is of order $n \times n$ but its rank is only $n - k$ (Appendix A.8).

The observable residuals e provide information about the unobservable disturbances u, but this information is not complete since k 'degrees of freedom' have been lost in the estimation process.

2.12 Estimation of σ^2

The linear model contains mean parameters β but also a variance parameter σ^2. We have seen how to estimate β. Let us now consider how to estimate σ^2, again applying the principle of best unbiased estimation.

A variance is a quadratic concept and therefore it makes little sense to construct a linear estimator of σ^2. But we can try to find a quadratic estimator, say $s^2 = y'By$, where we let B be positive semidefinite since s^2 must be nonnegative. We write s^2 as

$$\begin{aligned} s^2 &= y'By = (X\beta + u)'B(X\beta + u) \\ &= u'Bu + 2\beta'X'Bu + \beta'X'BX\beta, \end{aligned}$$

and take expectations. This gives

$$\mathrm{E}(s^2) = \mathrm{E}(u'Bu) + \beta'X'BX\beta = \sigma^2 \operatorname{tr}(B) + \beta'X'BX\beta,$$

since

$$\mathrm{E}(u'Bu) = \mathrm{E}\operatorname{tr}(Buu') = \operatorname{tr}\mathrm{E}(Buu') = \operatorname{tr}(B(\sigma^2 I_n)) = \sigma^2 \operatorname{tr}(B).$$

Unbiasedness of s^2 requires $\operatorname{tr}(B) = 1$ and $X'BX = 0$ (since $X'BX$ is symmetric, see Section A.5). The latter condition is equivalent to $BX = 0$ since B is positive semidefinite, see again Section A.5. Now, we know one positive semidefinite matrix which satisfies $BX = 0$, namely M. Its trace is $\operatorname{tr}(M) = n - k$; see Appendix A.8. Hence, if we choose $B = M/(n-k)$, then both constraints are satisfied, so that

$$s^2 = \frac{y'My}{n-k} = \frac{e'e}{n-k} \tag{2.6}$$

is an unbiased estimator of σ^2.

Of course, the matrix M is not the only positive semidefinite matrix satisfying $BX = 0$. We know that s^2 is unbiased but we don't know anything yet about other desirable properties. If, however, we add the assumption that the disturbances are normally distributed with mean zero and variance $\sigma^2 I_n$, then one can show (but we won't) that s^2 is best in the class of quadratic unbiased estimators in the sense that it achieves the lowest variance in that class.

Thus, while in the case of best unbiased estimation of β we added the restriction that the estimator is linear, here we need to add the restriction that the estimator is quadratic and also normality of y in order to obtain the best estimator; see also the discussion in Section 4.7.

2.13 Prediction

Suppose that we have estimated the linear model $y = X\beta + u$ given observations (y, X). Now we are given an $m \times k$ matrix X_* and we want to know how to 'predict' the corresponding value for y. (Prediction has nothing to do with the future, forecasting does.)

In order to answer this question we shall assume that the standard linear model is also valid for these new observations, so that

$$y_* = X_*\beta + u_*,$$

where u_* has mean zero, variance $\sigma^2 I_m$, and is uncorrelated with u. Our predictor will be

$$\hat{y}_* = X_*\hat{\beta} = X_*(X'X)^{-1}X'y, \tag{2.7}$$

in accordance with (1.4), and we want to know its properties. Of course, we have

$$\mathrm{E}(\hat{y}_*) = X_*\beta, \qquad \mathrm{var}(\hat{y}_*) = \sigma^2 X_*(X'X)^{-1}X_*'.$$

There are two ways to approach the prediction problem: we could be interested in $\mathrm{E}(y_*) = X_*\beta$ or in y_* itself.

If we are interested in $\mathrm{E}(y_*) = X_*\beta$, then Proposition 2.2 implies that $\hat{y}_* = X_*\hat{\beta}$ is the best linear unbiased estimator of $X_*\beta$.

If we are interested in y_* itself, then we should first realize that y_* is a random vector. Hence, strictly speaking, we cannot estimate it, because estimation is something that applies to nonrandom quantities, called parameters, not to random variables. We can, however, *predict* y_* as follows.

Define the *prediction error*

$$\hat{y}_* - y_* = X_*(\hat{\beta} - \beta) - u_* = X_*(X'X)^{-1}X'u - u_*.$$

The prediction error has mean zero and variance

$$\mathrm{var}(\hat{y}_* - y_*) = \sigma^2 \left(X_*(X'X)^{-1}X_*' + I_m\right).$$

Let $A'y$ be another linear predictor of y_*. Its prediction error is

$$A'y - y_* = (A'X - X_*)\beta + A'u - u_*.$$

Now define, as in Section 2.10,

$$D = A - X(X'X)^{-1}X_*'.$$

Then,

$$A'y - y_* = D'X\beta + (D' + X_*(X'X)^{-1}X')u - u_*.$$

The predictor is said to be *unbiased* if the prediction error has mean zero for all β, and this is the case if and only if $D'X = 0$. Under this restriction, the prediction error variance is

$$\operatorname{var}(A'y - y_*) = \sigma^2 \left(D'D + X_*(X'X)^{-1}X_*'\right) + I_m\big),$$

which is minimized when $D'D = 0$, that is, when $D = 0$. This shows that $A = X(X'X)^{-1}X_*'$ and hence that

$$A'y = X_*(X'X)^{-1}X'y = X_*\hat{\beta}$$

is the best linear unbiased predictor of y_*.

2.14 Restricted versus unrestricted model

In Section 1.9 we considered the two models

$$y = X_1\beta_1 + X_2\beta_2 + u, \tag{2.8}$$

which we shall now call the 'unrestricted' model, and

$$y = X_1\beta_1 + u_1, \tag{2.9}$$

which we shall call the 'restricted' model, where the restriction is $\beta_2 = 0$. We assume that the data are generated by (2.8), so that the unrestricted model coincides with the DGP and the smaller restricted model is underspecified.

The estimator of β_1 in the restricted model is of course

$$\hat{\beta}_{1r} = (X_1'X_1)^{-1}X_1'y, \tag{2.10}$$

while the estimator of β in the unrestricted model is given by

$$\hat{\beta}_u = \begin{pmatrix} \hat{\beta}_{1u} \\ \hat{\beta}_{2u} \end{pmatrix} = (X'X)^{-1}X'y,$$

where we can write the two subvectors as

$$\begin{pmatrix} \hat{\beta}_{1u} \\ \hat{\beta}_{2u} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_{1r} - (X_1'X_1)^{-1}X_1'X_2\hat{\beta}_{2u} \\ (X_2'M_1X_2)^{-1}X_2'M_1y \end{pmatrix}, \tag{2.11}$$

in view of (A.4). In the restricted model we find the mean and variance as

$$\mathrm{E}(\hat{\beta}_{1r}) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2, \qquad \mathrm{var}(\hat{\beta}_{1r}) = \sigma^2(X_1'X_1)^{-1}, \tag{2.12}$$

while in the unrestricted model the two estimators are unbiased,

$$\mathrm{E}(\hat{\beta}_{1u}) = \beta_1, \qquad \mathrm{E}(\hat{\beta}_{2u}) = \beta_2,$$

and their variances are given by

$$\mathrm{var}(\hat{\beta}_{1u}) = \sigma^2\left((X_1'X_1)^{-1} + \Delta\right) \tag{2.13}$$

and

$$\mathrm{var}(\hat{\beta}_{2u}) = \sigma^2(X_2'M_1X_2)^{-1}, \tag{2.14}$$

respectively, where

$$\Delta = (X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1(X_1'X_1)^{-1}.$$

The two estimators $\hat{\beta}_{1r}$ and $\hat{\beta}_{1u}$ of β_1 are always correlated, while $\hat{\beta}_{1u}$ and $\hat{\beta}_{2u}$ are only uncorrelated when $X_1'X_2 = 0$. In contrast, $\hat{\beta}_{1r}$ and $\hat{\beta}_{2u}$ are always uncorrelated. This follows because $M_1X_1 = 0$ and hence $\mathrm{cov}(X_1'y, X_2'M_1y) = \sigma^2X_1'M_1X_2 = 0$.

2.15 Mean squared error comparisons

In the previous section we obtained the first two moments of the restricted and the unrestricted estimators. This gives us two estimators of β_1, but how can we compare them?

If we compare two unbiased estimators, then we prefer the estimator with the lowest variance. But if we compare two estimators of which at least one is biased, then this is not a good strategy. Suppose for example that we wish to estimate θ. We have an unbiased estimator $\hat{\theta}_1$ and also a biased estimator $\hat{\theta}_2$. Let's take $\hat{\theta}_2 = 0$, a rather silly estimator, but simple. This estimator is biased (unless θ happens to be zero) but its variance is small, in fact zero. In such cases we can compare the mean squared errors, which take both bias and variance into account; see Section B.2. In the current example we have $\mathrm{MSE}(\hat{\theta}_1) = \mathrm{var}(\hat{\theta}_1)$ and $\mathrm{MSE}(\hat{\theta}_2) = \theta^2$, so that

$$\mathrm{MSE}(\hat{\theta}_1) \leq \mathrm{MSE}(\hat{\theta}_2) \iff \mathrm{var}(\hat{\theta}_1) \leq \theta^2.$$

This means that we prefer $\hat{\theta}_1$ unless the parameter θ is close to zero; then we prefer $\hat{\theta}_2 = 0$.

Let us now obtain the bias and the mean squared error of the two estimators of β_1 and compare them. The unrestricted estimator $\hat{\beta}_{1u}$ is unbiased, but the restricted estimator $\hat{\beta}_{1r}$ is biased and its bias is

$$\text{bias}(\hat{\beta}_{1r}) = \text{E}(\hat{\beta}_{1r} - \beta_1) = (X_1'X_1)^{-1}X_1'X_2\beta_2. \tag{2.15}$$

The mean squared errors of these two estimators are given by

$$\begin{aligned} V_r = \text{MSE}(\hat{\beta}_{1r}) &= \text{var}(\hat{\beta}_{1r}) + \Big(\text{bias}(\hat{\beta}_{1r})\Big)\Big(\text{bias}(\hat{\beta}_{1r})\Big)' \\ &= \sigma^2(X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'X_2\beta_2\beta_2'X_2'X_1(X_1'X_1)^{-1} \end{aligned}$$

and

$$V_u = \text{MSE}(\hat{\beta}_{1u}) = \text{var}(\hat{\beta}_{1u}) = \sigma^2(X_1'X_1)^{-1} + \sigma^2\Delta,$$

so that

$$V_r \geq V_u \iff X_1'X_2\beta_2\beta_2'X_2'X_1 \geq \sigma^2 X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1. \tag{2.16}$$

The condition in (2.16) is somewhat complicated. It is true that

$$\frac{\beta_2'X_2'M_1X_2\beta_2}{\sigma^2} \leq 1 \implies \beta_2\beta_2' \leq \sigma^2(X_2'M_1X_2)^{-1} \implies V_r \leq V_u,$$

but this condition is only sufficient, not necessary. However, in the special case $k_1 = k - 1$ and $k_2 = 1$ the condition is both necessary and sufficient and we find

$$V_r \geq V_u \iff \beta_2^2 X_1'x_2x_2'X_1 \geq \frac{\sigma^2}{x_2'M_1x_2}X_1'x_2x_2'X_1,$$

and hence, assuming that $X_1'x_2 \neq 0$,

$$V_r \geq V_u \iff \frac{|\beta_2|}{\sqrt{\sigma^2/x_2'M_1x_2}} \geq 1. \tag{2.17}$$

Let's try to understand this result. If $|\beta_2|$ is large then X_2 is apparently an important regressor and we should include it in the model. On the other hand, if $\text{var}(\hat{\beta}_{2u}) = \sigma^2/x_2'M_1x_2$ is large, then we don't learn much by including β_2 because we can't estimate it very precisely. So we want to include X_2 if either $|\beta_2|$ is large or $\sigma^2/x_2'M_1x_2$ is small or both.

2.16 Significance and importance

The previous discussion also provides an answer to the following puzzle. Suppose you are an econometrician working on a problem and some famous expert comes by, looks over your shoulder, and tells you that she *knows* the data-generation process. Of course, you yourself don't know the DGP. You use models but you don't know the truth; this expert does. Not only does the expert know the DGP but she is also willing to tell you, that is, she tells you the specification, not the actual parameter values. So now, you actually have the true model. What next? Is this the model that you are going to estimate?

The answer, surprisingly perhaps, is no. The truth, in general, is complex and contains many parameters, nonlinearities, and so on. All of these need to be estimated and this will produce large standard errors. There will be no bias if our model happens to coincide with the truth, but there will be large standard errors. A smaller model will have biased estimates but also smaller standard errors. Now, if we have a parameter in the true model whose value is small (so that the associated regressor is unimportant), then setting this parameter to zero will cause a small bias, because the size of the bias depends on the size of the deleted parameter:

$$\text{bias}(\hat{\beta}_{1r}) = (X_1'X_1)^{-1}X_1'X_2\beta_2,$$

according to (2.15). Setting this unimportant parameter to zero also means that we don't have to estimate it. The variance of the parameters of interest will therefore decrease, because

$$\begin{aligned}&\text{var}(\hat{\beta}_{1u}) - \text{var}(\hat{\beta}_{1r})\\ &\quad = \sigma^2(X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1(X_1'X_1)^{-1} \geq 0,\end{aligned}$$

using (2.12) and (2.14), and this decrease does *not* depend on the size of the deleted parameter. Thus, deleting a small unimportant parameter from the model is generally a good idea, because we will incur a small bias but may gain much precision.

This is true even if the estimated parameter happens to be highly 'significant', that is, have a large t-ratio, anticipating a concept that will be introduced in Section 3.5. Significance indicates that we have managed to estimate the parameter rather precisely, possibly because we have many observations. It does not mean that the parameter is important; see the discussion in Section 3.15.

We should therefore omit from the model all aspects that have little impact, so that we end up with a small model — one, which

captures the essence of our problem. As Einstein puts it: 'As simple as possible, but not simpler'.

2.17 Ottolenghi's ratatouille

I make a rather good ratatouille, based on a recipe from star chef Yotam Ottolenghi. Occasionally some friends drop by for dinner, usually unannounced, but expecting my ratatouille. Typically, I have some, but not all, ingredients at hand, so I have to improvise.

On a recent visit of my friends I had in stock: onions, garlic cloves, tomatoes, tomato purée, courgette, fresh coriander, rice, sunflower oil, caster sugar, and salt and black pepper. But I did not have the other required ingredients: fresh green chilli, red peppers, butternut squash, parsnip, French beans, aubergine, and potatoes. So I made the dish with what I had. Actually, the ratatouille tasted all right, although not quite as tasty as it is supposed to be.

A week later my friends dropped by again, and this time I had, in addition to all ingredients from last week, fresh green chilli and red peppers. Still not complete, but more complete than a week ago. Strangely, the ratatouille did not taste as good as one week ago.

This created a puzzle and a debate. How is it possible that getting closer to the true ingredients does not get us closer to the true taste? Of course, adding all ingredients creates the true taste as intended by Ottolenghi, but it seems that adding only some of them may not lead to an improvement. An addition in itself is not necessarily an improvement, it must be a 'balanced' addition.

2.18 Balanced addition

Let's continue our discussion on adding and omitting variables. The results in Sections 2.14–2.16 show what happens when we omit relevant variables. On the one hand we get biased estimators (which is bad), on the other hand the variance decreases (which is good), since $\text{var}(\hat{\beta}_{1r}) \leq \text{var}(\hat{\beta}_{1u})$.

However, this conclusion is only true when we compare the restricted model with an unrestricted model which coincides with the DGP. If, which is much more likely, we compare two models one of which is small (the restricted model) and the other is somewhat larger (the unrestricted model), but both are smaller than the DGP, then the estimator from the unrestricted model is also biased and, in fact, this bias may be *larger* than the bias from the restricted model.

This is what happened in the two partial executions of Ottolenghi's ratatouille.

Adding variables does *not* necessarily decrease the bias. The addition must be 'balanced'. But what is balanced? Since we don't know the DGP it is not easy to know whether an addition will be balanced or not. The application of econometrics requires more than mastering a collection of tricks. It also requires insight, intuition, and common sense.

Note to the reader:

This preview contains only the first two chapters. Chapters 3–6 and the two appendices are not available as preview.