Econometric Analysis of Cross Section and Panel Data

Second Edition

Jeffrey M. Wooldridge

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times Roman by Asco Typesetters, Hong Kong. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Wooldridge, Jeffrey M.
Econometric analysis of cross section and panel data / Jeffrey M. Wooldridge.—2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-262-23258-6 (hardcover : alk. paper)
1. Econometrics—Asymptotic theory. I. Title.
HB139.W663 2010
330.01′5195—dc22 2010020912

10 9 8 7 6 5 4 3 2 1

Preface

It has been almost 10 years since the first edition of *Econometric Analysis of Cross Section and Panel Data* was published. The reaction to the first edition was more positive than I could have imagined when I began thinking about the project in the mid-1990s. Of course, as several of you have kindly and constructively pointed out—and as was evident to me the first time I taught out of the book—the first edition was hardly perfect. Issues of organization and gaps in coverage were shortcomings that I wanted to address in a second edition from early on. Plus, there have been some important developments in econometrics that can and should be taught to graduate students in economics.

I doubt this second edition is perfect, either. But I believe it improves the first edition in substantive ways. The structure of this edition is similar to the first edition, but I have made some changes that will contribute to the reader's understanding of several topics. For example, Chapter 11, which covers more advanced topics in linear panel data models, has been rearranged to progress more naturally through situations where instrumental variables are needed in conjuction with methods for accounting for unobserved heterogeneity. Data problems—including censoring, sample selection, attrition, and stratified sampling—are now postponed until Chapters 19 and 20, after popular nonlinear models are presented under random sampling. I think this change will further emphasize a point I tried to make in the first edition: It is critical to distinguish between specifying a population model on the one hand and the method used to sample the data on the other. As an example, consider the Tobit model. In the first edition, I presented the Tobit model as applying to two separate cases: (1) a response variable is a corner solution outcome in the population (with the corner usually at zero) and (2) the underlying variable in the population is continuously distributed but the data collection scheme involves censoring the response in some way. Many readers commented that they were happy I made this distinction, because empirical researchers often seemed to confuse a corner solution due to economic behavior and a corner that is arbitrarily created by a data censoring mechanism. Nevertheless, I still found that beginners did not always fully appreciate the difference, and poor practice in interpreting estimates lingered. Plus, combining the two types of applications of so-called "censored regression models" gave short shrift to true data censoring. In this edition, models for corner solutions in the population are treated in Chapter 17, and a variety of data censoring schemes are covered in more detail in Chapter 19.

As in the first edition, I use the approach of specifying a population model and imposing assumptions on that model. Until Chapter 19, random sampling is assumed to generate the data. Unlike traditional treatments of, say, the linear regression model, my approach forces the student to specify the population of interest, propose

a model and assumptions in the population, and then worry about data issues. The last part is easy under random sampling, and so students can focus on various models that are used for populations with different features. The students gain a clear understanding that, under random sampling, our ability to identify parameters (and other quantities of interest) is a product of our assumed model in the population. Later it becomes clear that sampling schemes that depart from random sampling can introduce complications for learning about the underlying population.

The second edition continues to omit some important topics not covered in the first edition. The leading ones are simulation methods of estimation and semiparametric/nonparametric estimation. The book by Cameron and Trivedi (2005) does an admirable job providing accessible introductions to these topics.

I have added several new problems to each of the chapters. As in the first edition, the problems are a blend of methodological questions—some of which lead to tweaking existing methods in useful directions—and empirical work. Several data sets have been added to further illustate how more advanced methods can be applied. The data sets can be accessed by visiting links at the MIT Press website for the book: http://mitpress.mit.edu/9780262232586.

New to the Second Edition

Earlier I mentioned that I have reorganized some of the material from the first edition. I have also added new material, and expanded on some of the existing topics. For example, Chapter 6 (in Part II) introduces control function methods in the context of models linear in parameters, including random coefficient models, and discusses when the method is the same as two-stage least squares and when it differs. Control function methods can be used for certain systems of equations (Chapter 9) and are used regularly for nonlinear models to deal with endogenous explanatory variables, or heterogeneity, or both (Part IV). The control function method is convenient for testing whether certain variables are endogenous, and more tests are included throughout the book. (Examples include Chapter 15 for binary response models and Chapter 18 for count data.) Chapter 6 also contains a more detailed discussion of difference-in-differences methods for independently pooled cross sections.

Chapter 7 now introduces all of the different concepts of exogeneity of the explanatory variables in the context of panel data models, without explicitly introducing unobserved heterogeneity. This chapter also contains a detailed discussion of the properties of generalized least squares when an incorrect variance-covariance structure is imposed. This general discussion is applied in Chapter 10 to models that nominally impose a random effects structure on the variance-covariance matrix.

In this edition, Chapter 8 explicitly introduces and analyzes the so-called ''generalized instrumental variables'' (GIV) estimator. This estimator, used implicitly in parts of the first edition, is important for discussing efficient estimation. Further, some of the instrumental variables estimators used for panel data models in Chapter 11 are GIV estimators. It is helpful for the reader to understand the general idea underlying GIV, and to see its application to classes of important models.

Chapter 10, while focusing on traditional estimation methods for unobserved effects panel data models, demonstrates more clearly the relationships among random effects, fixed effects, and ''correlated random effects'' (CRE) models. While the first edition used the CRE approach often—especially for nonlinear models—I never used the phrase ''correlated random effects,'' which I got from Cameron and Trivedi (2005). Chapter 10 also provides a detailed treatment of the Hausman test for comparing the random and fixed effects estimators, including demonstrating that the traditional way of counting degrees of freedom when aggregate time effects are included is wrong and can be very misleading. The important topic of approximating the bias from fixed effects estimation and first differencing estimation, as a function of the number of available time periods, is also fleshed out.

Of the eight chapters in Part II, Chapter 11 has been changed the most. The random effects and fixed effects instrumental variables estimators are introduced and studied in some detail. These estimators form the basis for estimation of panel data models with heterogeneity and endogeneity, such as simultaneous equations models or models with measurement error, as well as models with additional orthogonality restrictions, such as Hausman and Taylor models. The method of first differencing followed by instrumental variables is also given separate treatment. This widely adopted approach can be used to estimate static models with endogeneity and dynamic models, such as those studied by Arellano and Bond (1991). The Arellano and Bond approach, along with several extensions, are now discussed in Section 11.6. Section 11.7 extends the treatment of models with individual-specific slopes, including an analysis of when traditional estimators are consistent for the population averaged effect, and new tests for individual-specific slopes.

As in the first edition, Part III of the book is the most technical, and covers general approaches to estimation. Chapter 12 contains several important additions. There is a new discussion concerning inference when the first-step estimation of a two-step procedure is ignored. Resampling schemes, such as the bootstrap, are discussed in more detail, including how one used the bootstrap in microeconometric applications with a large cross section and relatively few time periods. The most substantive additions are in Sections 12.9 and 12.10, which cover multivariate nonlinear least squares and quantile methods, respectively. An important feature of Section 12.9 is

that I make a simple link between mutivariate weighted nonlinear least squares—an estimation method familiar to economists—and the generalized estimating equations (GEE) approach. In effect, these approaches are the same, a point that hopefully allows economists to read other literature that uses the GEE nomenclature.

The section on quantile estimation covers different asymptotic variance estimators and discusses how they compare to violation of assumptions in terms of robustness. New material on estimating and inference when quantile regression is applied to panel data gives researchers simple methods for allowing unobserved effects in quantile estimation, while at the same time offering inference that is fully robust to arbitrary serial correlation.

Chapter 13, on maximum likelihood methods, also includes several additions, including a general discussion of nonlinear unobserved effects models and the different approaches to accounting for the heterogeneity (broadly, random effects, "fixed" effects, and correlated random effects) and different estimation methods (partial maximum likelihood or full maximum likelihood). Two-step maximum likelihood estimators are covered in more detail, including the case where estimating parameters in a first stage can be more efficient than simply plugging in known population values in the second stage. Section 13.11 includes new material on quasi-maximum likelihood estimation (QMLE). This section argues that, for general misspecification, only one form of asymptotic variance can be used. The QMLE perspective is attractive in that it admits that models are almost certainly wrong, thus we should conduct inference on the approximation in a valid way. Vuong's (1988) model selection tests, for nonnested models, is explicitly treated as a way to choose among competing models that are allowed to be misspecified. I show how to extend Vuong's approach to panel data applications (as usual, with a relatively small number of time periods).

Chapter 13 also includes a discussion of QMLE in the linear exponential family (LEF) of likelihoods, when the conditional mean is the object of interest. A general treatment allows me to appeal to the consistency results, and the methods for inference, at several points in Part IV. I emphasize the link between QMLE in the LEF and the so-called "generalized linear models" (GLM) framework. It turns out that GLM is just a special case of QMLE in the LEF, and this recognition should be helpful for studying research conducted from the GLM perspective. A related topic is the GEE approach to estimating panel data models. The starting point for GEE in panel data is to use (for a generic time period) a likelihood in the LEF, but to regain some efficiency that has been lost by not implementing full maximum likelihood by using a generalized least squares approach.

Chapter 14, on generalized method of moments (GMM) and minimum distance (MD) estimation, has been slightly reorganized so that the panel data applications

come at the end. These applications have also been expanded to include unobserved effects models with time-varying loads on the heterogeneity.

Perhaps for most readers the changes to Part IV will be most noticeable. The material on discrete response models has been split into two chapters (in contrast to the rather unwieldy single chapter in the first edition). Because Chapter 15 is the first applications-oriented chapter for nonlinear models, I spend more time discussing different ways of measuring the magnitudes of the effects on the response probability. The two leading choices, the partial effects evaluated at the averages and the average partial effect, are discussed in some detail. This discussion carries over for panel data models, too. A new subsection on unobserved effects panel data models with unobserved heterogeneity and a continuous endogenous explanatory variable shows how one can handle both problems in nonlinear models. This chapter contains many more empirical examples than the first edition.

Chapter 16 is new, and covers multinomial and ordered responses. These models are now treated in more detail than in the first edition. In particular, specification issues are fleshed out and the issues of endogeneity and unobserved heterogeneity (in panel data) are now covered in some detail.

Chapter 17, which was essentially Chapter 16 in the first edition, has been given a new title, Corner Solutions Responses, to reflect its focus. In reading Tobin's (1958) paper, I was struck by how he really was talking about the corner solution case—data censoring had nothing to do with his analysis. Thus, this chapter returns to the roots of the Tobit model, and covers several extensions. An important addition is a more extensive treatment of two-part models, which is now in Section 17.6. Hopefully, my unified approach in this section will help clarify the relationships among so-called "hurdle" and "selection" models, and show that the latter are not necessarily superior. Like Chapter 15, this chapter contains several more empirical applications.

Chapter 18 covers other kinds of limited dependent variables, particularly count (nonnegative integer) outcomes and fractional responses. Recent work on panel data methods for fractional responses has been incorporated into this chapter.

Chapter 19 is an amalgamation of material from several chapters in the first edition. The theme of Chapter 19 is data problems. The problem of data censoring—where a random sample of units is obtained from the population, but the response variable is censored in some way—is given a more in-depth treatment. The extreme case of binary censoring is included, along with interval censoring and top coding. Readers are shown how to allow for endogenous explanatory variables and unobserved heterogeneity in panel data.

Chapter 19 also includes the problem of not sampling at all from part of the population (truncated sampling) or not having any information about a response for a

subset of the population (incidental truncation). The material on unbalanced panel data sets and the problems of incidental truncation and attrition in panel data are studied in more detail, including the method of inverse probability weighting for correcting for missing data.

Chapter 20 continues the material on nonrandom sampling, providing a separate chapter for stratified sampling and cluster sampling. Stratification and clustering are often features of survey data sets, and it is important to know what adjustments are required to standard econometric methods. The material on cluster sampling summarizes recent work on clustering with a small number of clusters.

The material on treatment effect estimation is now in Chapter 21. While I preserved the setup from the first edition, I have added several more topics. First, I have expanded the discussion of matching estimators. Regression discontinuity designs are covered in a separate section.

The final chapter, Chapter 22, now includes the introductory material on duration analysis. I have included more empirical examples than were in the first edition.

Possible Course Outlines

At Michigan State, I teach a two-semester course to second-year, and some third-year, students that covers the material in my book—plus some additional material. I assume that the graduate students know, or will study on their own, material from Chapters 2 and 3. It helps move my courses along when students are comfortable with the basic algebra of probability (conditional expectations, conditional variances, and linear projections) as well as the basic limit theorems and manipulations. I typically spend a few lectures on Chapters 4, 5, and 6, primarily to provide a bridge between a more traditional treatment of the linear model and one that focuses on a linear population model under random sampling. Chapter 6 introduces control function methods in a simple context and so is worth spending some time on.

In the first semester (15 weeks), I cover the material (selectively) through Chapter 17. But I currently skip, in the first semester, the material in Chapter 12 on multivariate nonlinear regression and quantile estimation. Plus, I do not cover the asymptotic theory underlying M-estimation in much detail, and I pretty much skip Chapter 14 altogether. In effect, the first semester covers the popular linear and nonlinear models, for both cross section and panel data, in the context of random sampling, providing much of the background needed to justify the large-sample approximations.

In the second semester I return to Chapter 12 and cover quantile estimation. I also cover the general quasi-MLE and generalized estimating equations material in

Chapter 13. In Chapter 14, I find the minimum distance approach to estimation is important as a more advanced estimation method. I cover some of the panel data examples from this chapter. I then jump to Chapter 18, which covers count and fractional responses. I spend a fair amount of time on Chapters 19 and 20 because data problems are especially important in practice, and it is important to understand the strengths and weakness of the competing methods. After I cover the main parts of Chapter 21 (including regression discontinuity designs) and Chapter 22 (duration analysis), I sometimes have extra time. (However, if I were to cover some of the more advanced topics in Chapter 21—multivalued and multiple treatments, and dynamic treatement effects in the context of panel data—I likely would run out of time.) If I do have extra time, I like to provide an introduction to nonparametric and semiparametric methods. Cameron and Trivedi (2005) is accessible for the basic methods, while the book by Li and Racine (2007) is comprehensive. Illustrating nonparametric methods using the treatment effects material in Chapter 21 seems particularly effective.

Supplements

A student *Solutions Manual* is available that includes answers to the odd-numbered problems (see http://mitpress.mit.edu/9780262731836). Any instructor who adopts the book for a course may have access to all solutions. In addition, I have created a set of slides for the two-semester course that I teach. They are available as Scientific Word 5.5 files—which can be edited—or as pdf files. For these teaching aids see the web page for the second edition: http://mitpress.mit.edu/9780262232586.