Preface

The first edition of *An Introduction to Modern Econometrics Using Stata* was published 20 years ago, with examples drawn from Stata 9. This second edition was written using StataNow 19.5 and highlights some of the important features that have been incorporated into Stata since Stata 9. The growth of Stata's own capabilities and the increasing amount of community-contributed commands has made Stata's role in the statistical software universe all the more prominent.

In retrospect, the most significant enhancement to Stata's ease of use was the introduction of factor-variable notation and, along with it, the `margins` command. More recently, the introduction of table production commands has addressed one of the key aspects of researchers' efficiency: the ability to produce publication-quality results in various formats. These features, combined with the computational improvements of embracing 64-bit processors, have played important roles in Stata's popularity in many disciplines, including my own fields of applied econometrics and public health.

What follows is drawn from the preface to the first edition of this book. The goals are largely unchanged, but many aspects of what an applied researcher needs to complete their tasks have been updated by the evolution of Stata's capabilities. Much of the earlier edition has been rewritten to reflect those improvements. I highlight recent changes, such as the introduction of frames and table production commands. I also present econometric methods such as the computation of predictive means and average marginal effects, the correlated random-effects model for panel data, and new methods for IV estimation. The last chapter, on discrete and limited dependent variables, has been expanded to include fractional response models and count-data models.

This book is a concise guide for applied researchers in economics, finance, and related disciplines to learn basic econometrics and work through Stata examples using typical datasets analyzed in economics. Readers should be familiar with applied statistics at the level of a simple linear regression (ordinary least-squares) model and its algebraic representation. Knowledge of these concepts would be equivalent to the level of an undergraduate statistics or econometrics course sequence such as that contained in Wooldridge (2025). The book also uses some multivariate calculus (partial derivatives) and linear algebra.

I presume that the reader is familiar with Stata's windowed interface and with the basics of data input, data transformation, and descriptive statistics. Readers who are comfortable interacting with Stata should feel free to skip to chapter 4, where the discussion of econometrics begins in earnest.

In any research project, much of the effort involves the preparation of the data specified as part of an econometric model. While this book primarily focuses on applied econometric practice, we must consider the challenges that many researchers face in moving from their original data sources to the form needed in an econometric model—or even the form needed to provide appropriate tabulations and graphs for the project. Accordingly, chapter 2 focuses on the details of data management and several tools available in Stata to ensure that the appropriate transformations are accomplished accurately and efficiently. If you are familiar with these aspects of Stata usage, then feel free to skim this material, perhaps returning to it to refresh your understanding of Stata usage. Likewise, chapter 3 discusses the organization of economic and financial data and the Stata commands needed to reorganize data among the several forms of organization. If you are eager to begin with the econometrics of linear regression, you need only skim this chapter, noting its content for future reference.

Chapter 4 begins the econometric content of this book and presents the most widely used tool for econometric analysis: the multiple linear regression model applied to continuous and categorical variables. The chapter also discusses how to interpret and present regression estimates and discusses the logic of hypothesis tests and linear and nonlinear restrictions. The last section of the chapter considers residuals, predicted values, and marginal effects.

Applying the regression model depends on some assumptions that real datasets often violate. Chapter 5 discusses how the crucial zero-conditional-mean assumption of the errors can be violated in the presence of specification errors. The chapter also discusses statistical and graphical techniques for detecting specification errors. Chapter 6 focuses on estimation of the variance–covariance matrix, discussing robust, cluster–robust, and autocorrelation-consistent structures, as well as the generalized linear regression model. It also explains how to diagnose and correct for heteroskedasticity and serial correlation. Chapter 7 discusses the usage of indicator or dummy variables in linear regression models containing both quantitative and qualitative factors, models with interaction effects, and models of structural change.

Many regression models in applied economics violate the zero-conditional-mean assumption of the errors because they simultaneously determine the response variable and one or more regressors or because of measurement error in the regressors. No matter the cause, ordinary least-squares techniques will no longer generate unbiased and consistent estimates, so you must use instrumental-variables (IV) techniques instead. Chapter 8 presents the IV estimator and its generalized method of moments counterpart along with tests for determining the need for IV techniques.

Chapter 9 applies models to panel or longitudinal data that have both cross-sectional and time-series dimensions. Extensions of the regression model allow you to take advantage of the rich information in panel data, accounting for the heterogeneity in both panel unit and time dimensions.

Many econometric applications model categorical and limited dependent variables: a binary outcome—such as a purchase decision, a count, or a constrained response such

as the amount spent—that combines the decision whether to purchase with the decision of how much to spend, conditional on purchasing. Because linear regression techniques are generally not appropriate for modeling these outcomes, chapter 10 presents several limited dependent variable estimators available in Stata.

The appendixes discuss techniques for importing external data into Stata and explain basic Stata programming. Although you can use Stata without doing any programming, learning how to program in Stata can help you save a lot of time and effort. You should also learn to generate reproducible results by using do-files that you can document, archive, and rerun. Following Stata's guidelines will make your do-files shorter and easier to maintain and modify.

Brighton, MA
Oak Square School
September 2026

Christopher F. Baum