An Introduction to Modern Econometrics Using Stata

Second Edition

CHRISTOPHER F. BAUM
Department of Economics and School of Social Work
Boston College



A Stata Press Publication
StataCorp LLC
College Station, Texas



Copyright © 2006, 2027 by StataCorp LLC
All rights reserved. First edition 2006
Second edition 2027

Published by Stata Press, 4905 Lakeway Drive, College Station, Texas 77845
Typeset in LaTeX 2$_\varepsilon$
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

Print ISBN-10: 1-59718-457-8
Print ISBN-13: 978-1-59718-457-1
ePub ISBN-10: 1-59718-458-6
ePub ISBN-13: 978-1-59718-458-8

Library of Congress Control Number: 2026950925

No part of this book may be reproduced, stored in a retrieval system, or transcribed, in any form or by any means—electronic, mechanical, photocopy, recording, or otherwise—without the prior written permission of StataCorp LLC.

Stata, STATA, Stata Press, Mata, MATA, NetCourse, and NetCourseNow are registered trademarks of StataCorp LLC.

Stata and Stata Press are registered trademarks with the World Intellectual Property Organization of the United Nations.

StataNow is a trademark of StataCorp LLC.

LaTeX 2$_\varepsilon$ is a trademark of the American Mathematical Society.

Other brand and product names are registered trademarks or trademarks of their respective companies.

Contents

List of figures **xv**

List of tables **xvii**

Preface **xix**

Acknowledgments **xxiii**

Notation and typography **xxv**

1 Introduction **1**

1.1 An overview of Stata's distinctive features 1

1.2 Installing the necessary software 4

1.3 Installing the support materials 5

2 Working with economic and financial data in Stata **7**

2.1 The Stata workspace 7

2.1.1 Accessing a Stata dataset 7

2.1.2 Data types 9

2.1.3 _n and _N 11

2.1.4 The varlist 11

2.1.5 The numlist 11

2.1.6 Stata's fundamental design 11

2.1.7 if exp and in range 12

2.1.8 sort and gsort 14

2.1.9 Command syntax 15

2.1.10 return list, ereturn list 15

2.2 Modifying the workspace 17

2.2.1 generate and replace 17

Logical operators 18

2.2.2 egen . . . 18
2.2.3 label and notes . . . 21
2.2.4 drop and keep . . . 24
2.2.5 rename . . . 24
2.2.6 save, saveold, and compress . . . 24
2.3 Factor variables . . . 25
2.3.1 The interaction operator . . . 25
2.3.2 The factorial interaction operator . . . 26
2.4 Prefix operators . . . 26
2.4.1 The by prefix . . . 26
2.4.2 Computation for by-groups . . . 28
2.4.3 The statsby prefix . . . 32
2.4.4 The rolling prefix . . . 33
2.5 Data management . . . 33
2.5.1 Data input . . . 33
2.5.2 Handling missing data . . . 34
2.5.3 String-to-numeric conversion and vice versa . . . 35
2.5.4 Recoding discrete and continuous variables . . . 37
2.5.5 Handling dates and times . . . 39
2.5.6 Some useful functions for generate or replace . . . 40
2.5.7 Data validation with assert, count, and duplicates . . . 42
2.6 Local and global macros, scalars, and matrices . . . 49
2.6.1 Local macros . . . 49
2.6.2 Loops with forvalues and foreach . . . 52
Nested loops . . . 54
2.6.3 Extended macro functions and list functions . . . 55
2.6.4 Global macros . . . 56
2.6.5 System parameters, settings and constants . . . 56
2.6.6 Scalars and matrices . . . 56
2.7 The Mata language . . . 58

2.8 Python integration with Stata 61
2.8.1 Julia integration with Stata 64
2.9 Adding capabilities to Stata 65
2.9.1 The SSC Archive 65
2.9.2 The Stata Journal 65
3 Organizing and handling economic data 67
3.1 Cross-sectional data and identifier variables 67
3.2 Time-series data 68
3.2.1 Time-series calendar 68
3.2.2 Time-series operators 69
3.3 Pooled cross-sectional time-series data 70
3.3.1 Combining cross-sectional and time-series datasets 70
3.3.2 Creating long-format datasets with append 71
3.3.3 Using merge to add aggregate characteristics 72
3.3.4 Using joinby rather than merge 73
3.3.5 Other useful data management commands 75
fillin 75
cross 76
stack 76
separate 77
3.4 Panel data 78
3.4.1 Operating on panel data 80
3.4.2 Unbalanced panels and data screening 82
3.4.3 Other transforms of panel data 85
3.4.4 Moving-window summary statistics and correlations 86
3.4.5 Reshaping the data 88
3.5 Working with frames 92
3.5.1 Working with linked datasets 92
3.6 Reporting descriptive statistics 101
3.6.1 dtable 102

3.6.2 collect . . . 104
3.6.3 table . . . 104
4 Linear regression 109
4.1 Introduction . . . 109
4.2 Computing linear regression estimates . . . 110
4.2.1 Regression as a method of moments estimator . . . 112
4.2.2 The sampling distribution of regression estimates . . . 113
4.2.3 Efficiency of the regression estimator . . . 114
4.2.4 Numerical identification of the regression estimates . . . 115
4.3 Interpreting regression estimates . . . 116
4.3.1 Research project: A study of wage determination . . . 116
4.3.2 The ANOVA table: ANOVA F and R^2 . . . 123
4.3.3 Adjusted R^2 . . . 126
4.3.4 The coefficient estimates and beta coefficients . . . 127
4.3.5 Regression without a constant term . . . 129
4.3.6 Detecting collinearity in regression . . . 130
4.4 Presenting regression estimates . . . 133
4.4.1 etable . . . 135
4.4.2 estout . . . 138
4.5 Hypothesis tests, linear restrictions, and constrained least squares . . . 139
4.5.1 Wald tests with test . . . 142
4.5.2 Wald tests involving linear combinations of parameters . . . 143
4.5.3 Joint hypothesis tests . . . 145
4.5.4 Testing nonlinear restrictions and forming nonlinear combinations . . . 147
4.5.5 Testing competing (nonnested) models . . . 148
4.6 Computing residuals and predicted values . . . 150
4.6.1 Computing interval predictions . . . 152
4.7 Computing predictive means and marginal effects . . . 156
4.7.1 Average marginal effects . . . 160

4.A Appendix: Regression as a least-squares estimator 167
4.B Appendix: The large-sample VCE for linear regression 168
5 Specifying the functional form 171
5.1 Introduction 171
5.2 Specification error 171
5.2.1 Omitting relevant variables from the model 172
5.2.2 Graphically analyzing regression data 173
5.2.3 Added-variable plots 174
5.2.4 Including irrelevant variables in the model 176
5.2.5 Model selection 176
Specifying dynamics in time-series regression models 177
5.2.6 Misspecification of the functional form 178
5.2.7 Ramsey's RESET 178
5.2.8 Specification plots 179
5.2.9 Specification and interaction terms 180
5.2.10 Outlier statistics and measures of leverage 182
The DFITS statistic 184
The DFBETA statistic 187
5.3 Choosing the functional form 190
5.4 Endogeneity and measurement error 194
5.4.1 Errors-in-variables 196
6 Variance–covariance matrix estimation 199
6.1 The generalized linear regression model 200
6.1.1 Types of deviations from the classical VCE 201
6.1.2 The robust estimator of the VCE 202
6.1.3 The HC2 and HC3 estimators of the VCE 203
6.1.4 The cluster–robust estimator of the VCE 205
6.1.5 The Newey–West estimator of the VCE 209
6.1.6 The generalized least-squares estimator 211
The FGLS estimator 212

6.2 Heteroskedasticity in the error distribution . . . 213
6.2.1 Heteroskedasticity related to scale . . . 214
Testing for heteroskedasticity related to scale . . . 214
FGLS estimation . . . 217
6.2.2 Heteroskedasticity between groups of observations . . . 219
Testing for heteroskedasticity between groups of observations 220
FGLS estimation . . . 221
6.2.3 Heteroskedasticity in grouped data . . . 222
FGLS estimation . . . 223
6.3 Serial correlation in the error distribution . . . 224
6.3.1 Testing for serial correlation . . . 225
6.3.2 FGLS estimation with serial correlation . . . 229
7 Regression with indicator variables 231
7.1 Testing for significance of a qualitative factor . . . 231
7.1.1 Regression with one qualitative measure . . . 232
7.1.2 Regression with two qualitative measures . . . 234
Interaction effects . . . 238
7.2 Regression with qualitative and quantitative factors . . . 239
Testing for slope differences . . . 241
7.3 Seasonal adjustment with indicator variables . . . 248
7.4 Testing for structural stability and structural change . . . 252
7.4.1 Constraints of continuity and differentiability . . . 253
7.4.2 Structural change in a time-series model . . . 258
8 Instrumental-variables estimators 263
8.1 Introduction . . . 263
8.2 Endogeneity in economic relationships . . . 263
8.3 2SLS . . . 266
8.4 ivregress . . . 267
8.5 Identification and tests of overidentifying restrictions . . . 268
8.6 Computing IV estimates . . . 270

8.7 GMM estimation and ivreg2 273
8.7.1 The GMM estimator 273
8.7.2 GMM in a homoskedastic context 275
8.7.3 GMM and heteroskedasticity-consistent standard errors . . 275
8.7.4 GMM and clustering 277
8.7.5 GMM and heteroskedasticity- and autocorrelation -consistent standard errors 278
8.8 Testing overidentifying restrictions in GMM 280
8.8.1 Testing a subset of the overidentifying restrictions in GMM 281
8.9 Testing for heteroskedasticity in the IV context 287
8.10 Testing for serial correlation in the IV context 289
8.11 Testing the relevance of instruments 292
8.12 Durbin–Wu–Hausman tests for endogeneity in IV estimation 297
8.13 Heteroskedasticity-based estimation 301
8.13.1 The model and estimator 302
8.13.2 Implementing the estimator and tests 303
8.13.3 Illustration of ivreg2h 305
8.14 Control-function estimation 311
8.A Appendix: Omitted-variables bias 314
8.B Appendix: Measurement error 314
8.B.1 Solving errors-in-variables problems 316
9 Panel-data models **319**
9.1 FE, RE, and CRE models 320
9.1.1 One-way FE 321
9.1.2 Time effects and two-way FE 324
9.1.3 The between estimator 326
9.1.4 One-way RE 327
9.1.5 Testing the appropriateness of RE 330
9.1.6 CRE 332
9.1.7 Prediction from FE, RE, and CRE models 335

9.1.8 FE models with multiple FE . . . 335
9.2 IV models for panel data . . . 337
9.3 DPD models . . . 338
9.4 SUR models . . . 346
9.4.1 SUR with identical regressors . . . 351
9.5 Moving-window regression estimates . . . 351
10 Models of discrete and limited dependent variables 355
10.1 Binomial logit and probit models . . . 355
10.1.1 The latent variable approach . . . 356
10.1.2 Marginal effects and predictions . . . 358
Binomial probit . . . 358
Binomial logit and grouped logit . . . 364
10.1.3 Evaluating specification and goodness of fit . . . 368
10.2 Fractional response models . . . 372
10.3 Ordered logit and probit models . . . 374
10.4 Models of count data . . . 377
10.4.1 The Poisson model . . . 378
10.4.2 The negative binomial model . . . 380
10.4.3 Extended count-data models . . . 382
10.5 Truncated regression and tobit models . . . 383
10.5.1 Truncation . . . 384
10.5.2 Censoring . . . 386
10.6 Incidental truncation and sample-selection models . . . 392
10.7 Bivariate probit and probit with selection . . . 396
10.7.1 Binomial probit with selection . . . 397
A Getting the data into Stata 401
A.1 Inputting data from ASCII text files and spreadsheets . . . 401
A.1.1 Handling text files . . . 402
Free format versus fixed format . . . 402
import delimited . . . 404

A.1.2 Accessing data stored in spreadsheets 405
A.1.3 Fixed-format data files 405
A.2 Importing data from other package formats 410
B The basics of Stata programming **413**
B.1 The program and syntax statements 414
B.2 Using Mata functions in Stata programs 420
References **427**
Author index **441**
Subject index **445**

(Pages omitted)

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

(Pages omitted)

4 Linear regression

This chapter discusses the most widely used tool in applied economics: the linear regression model, which relates a set of continuous variables to a continuous outcome. The explanatory variables in a regression model often include one or more binary or indicator variables; see chapter 7. Likewise, many models seek to explain a binary response variable as a function of a set of factors, which linear regression does not handle well. Chapter 10 discusses several forms of that model, including those in which the response variable is limited but not binary.

4.1 Introduction

This chapter discusses multiple regression in the context of a prototype economic research project. To carry out such a research project, we must

1. define a research framework—or economic model—that lets us specify the questions of interest and defines how we will interpret the empirical results;
2. find a dataset containing empirical counterparts to the quantities specified in the economic model;
3. use exploratory data analysis to familiarize ourselves with the data and identify outliers, extreme values, and the like;
4. fit the model and use specification analysis to determine the adequacy of the explanatory factors and their functional form;
5. conduct statistical inference (given satisfactory findings from specification analysis) on the research questions posed by the model; and
6. analyze the findings from hypothesis testing and the success of the model in terms of predictions and marginal effects. Based on these findings, we may have to return to one of the earlier stages to reevaluate the dataset and its specification and functional form.

Section 4.2 reviews the basic regression analysis theory on which regression point and interval estimates are based. Section 4.3 introduces a prototype economic research project studying the determinants of wage determination and discusses the various components of Stata's results from fitting a regression model of wages. Section 4.4 discusses how to transform Stata's estimation results into publication-quality tables. Section 4.5 discusses hypothesis testing and estimation subject to constraints on the parameters.

Section 4.6 deals with computing residuals and predicted values. The last section discusses computing marginal effects. In the following chapters, we take up violations of the assumptions on which regression estimates are based.

4.2 Computing linear regression estimates

The linear regression model is the most widely used econometric model and the baseline against which all others are compared. It specifies the conditional mean of a response variable y as a linear function of k independent variables

$$E(y \mid x_1, x_2, \ldots, x_k) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Given values for the βs, which are fixed parameters, the regression function predicts the average value of y in the population for different values of $x_1, x_2, \ldots, x_k$.

Suppose that the mean value of an individual's wage, conditional on their years of work experience, is given by

$$E(\texttt{wage} \mid \texttt{ttl_exp}) = \beta_1 + \beta_2\ \texttt{ttl_exp}$$

where `ttl_exp` is years of their work experience. This relationship reflects the hypothesis that wages will generally increase with human capital formation, captured by their experience in the labor market.

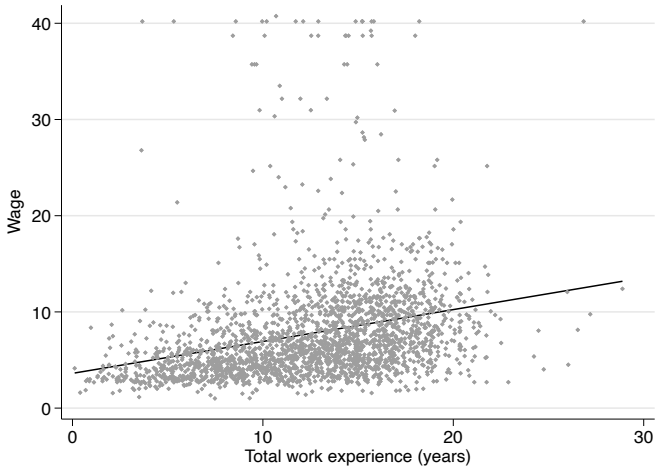


Figure 4.1. Conditional mean of wage

Figure 4.1 shows hourly wages recorded for 2,246 individuals, along with the linear fit of wages to years of work experience. The regression function's estimate of the conditional mean of `wage` for each value of `ttl_exp` is shown by the appropriate point

on the line. As theory predicts, wages rise with human capital formation. Of course, this relationship between the wage and years of experience must be considered ceteris paribus: all other factors that might affect the individual's wage are held constant when we evaluate the effect of additional years of work experience on his or her wage.

When we are working with economic data, we do not know the population values of $\beta_1, \beta_2, \ldots, \beta_k$. We work with a sample of N observations of data from that population. Using the information in this sample, we must

1. obtain consistent estimates of the coefficients $(\beta_1, \beta_2, \ldots, \beta_k)$,
2. determine how much our coefficient estimates would change if we were given another sample from the same population,
3. decide whether there is enough evidence to rule out some values for some of the coefficients $(\beta_1, \beta_2, \ldots, \beta_k)$, and
4. use our estimated $(\beta_1, \beta_2, \ldots, \beta_k)$ to interpret the model.

To obtain estimates of the coefficients, some assumptions must be made about the process that generated the data. I discuss those assumptions below and describe what I mean by good estimates. Before performing steps 2–4, I check whether the data support these assumptions by using a process known as specification analysis.

If we have a cross-sectional sample from the population, the linear regression function for each observation in the sample has the form

$$y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \cdots + \beta_k x_{i,k} + u_i$$

for each observation in the sample $i = 1, 2, \ldots, N$. The u process is a stochastic disturbance representing the net effect of all other unobservable factors that might influence y. The variance of its distribution, σ_u^2, is an unknown population parameter to be estimated along with the β parameters. We assume that $N > k$; to conduct statistical inference, there must be more observations in the sample than parameters to be estimated. In practice, N should be much larger than k.

We can write the linear regression function in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{4.1}$$

where $\mathbf{X}$ is an $(N \times k)$ matrix of sample values.[1]

This population regression function specifies that a set of k regressors in $\mathbf{X}$ and the stochastic disturbance u are the determinants of the response variable (or regressand) y. We usually assume that the regression function contains a constant term, so x_1 is understood to equal 1 for each observation.

1. Some textbooks use k in this context to refer to the number of slope parameters rather than the number of columns of $\mathbf{X}$. That will explain the deviations in the formulas given below; where I write k, some authors write $(k + 1)$.

The key assumption in the linear regression model involves the relationship in the population between the regressors $\mathbf{x}$ and u.[2] We may rewrite (4.1) as

$$u = y - \mathbf{x}\boldsymbol{\beta}$$

We assume that

$$E(u \mid \mathbf{x}) = 0 \tag{4.2}$$

that is, that the u process has a zero conditional mean. This assumption is that the unobserved factors involved in the regression function are not related systematically to the observed factors. This approach to linear regression lets us consider both nonstochastic and stochastic regressors in $\mathbf{X}$ without distinction as long as they satisfy assumption (4.2).[3]

We also assume conditional homoskedasticity of the error process:

$$E\left(u^2 \mid \mathbf{x}\right) = \sigma_u^2 \tag{4.3}$$

This assumption states that the error variance, conditional on the regressors, is constant. Given the zero-conditional-mean assumption, this also implies that the unconditional expectation of u^2 is σ_u^2. This assumption is testable, as we will see in our description of specification analysis. The assumption of conditional homoskedasticity also implies that

$$\operatorname{Var}(y \mid \mathbf{x}) = \sigma_u^2$$

4.2.1 Regression as a method of moments estimator

We can use the zero-conditional-mean assumption shown in (4.2) to define a method of moments estimator of the regression function. Method of moments estimators are defined by moment conditions that are assumed to hold for the population moments. When we replace the unobservable population moments by their sample counterparts, we derive feasible estimators of the model's parameters. The zero-conditional-mean assumption gives rise to a set of k moment conditions, one for each x. In particular, the zero-conditional-mean assumption implies that each regressor is uncorrelated with u.[4]

$$\begin{aligned} E(\mathbf{x}'u) &= \mathbf{0} \\ E\{\mathbf{x}'(y - \mathbf{x}\boldsymbol{\beta})\} &= \mathbf{0} \end{aligned} \tag{4.4}$$

2. $\mathbf{x}$ is a vector of random variables, and u is a scalar random variable. In (4.1), $\mathbf{X}$ is a matrix of realizations of the random vector $\mathbf{x}$, and $\mathbf{u}$ and $\mathbf{y}$ are vectors of realizations of the scalar random variables u and y.
3. Chapter 8 discusses how to use the instrumental-variables estimator when the zero-conditional-mean assumption is violated.
4. The assumption of a zero conditional mean is stronger than that of a zero covariance because covariance considers only linear relationships between the random variables.

Substituting calculated moments from our sample into the expression and replacing the unknown coefficients $\boldsymbol{\beta}$ with estimated values $\widehat{\boldsymbol{\beta}}$ in (4.4) yields the ordinary least-squares (OLS) estimator

$$\begin{aligned} \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} &= \mathbf{0} \\ \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned} \tag{4.5}$$

We can use $\widehat{\boldsymbol{\beta}}$ to calculate the regression residuals:

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$$

Given the solution for the vector $\widehat{\boldsymbol{\beta}}$, the additional parameter of the regression problem σ_u^2—the population variance of the stochastic disturbance—can be estimated as a function of the regression residuals $\widehat{u}_i$,

$$s^2 = \frac{\sum_{i=1}^{N} \widehat{u}_i^2}{N-k} = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{N-k} \tag{4.6}$$

where $(N-k)$ is the residual degrees of freedom of the regression problem. The positive square root of s^2 is often termed the standard error of regression, or root mean squared error. Stata uses the latter term and displays s as `Root MSE`.

The method of moments is not the only approach for deriving the linear regression estimator of (4.5), which is the well-known formula from which the OLS estimator is derived.[5]

4.2.2 The sampling distribution of regression estimates

The OLS estimator $\widehat{\boldsymbol{\beta}}$ is a vector of random variables because it is a function of the random variable y, which in turn is a function of the stochastic disturbance u. The OLS estimator takes on different values for each sample of N observations drawn from the population. Because we often have only one sample to work with, we may be unsure of the usefulness of the estimates from that sample. The estimates are the realizations of the random vector $\widehat{\boldsymbol{\beta}}$ from the sampling distribution of the OLS estimator. To evaluate the precision of a given vector of estimates $\widehat{\boldsymbol{\beta}}$, we use the sampling distribution of the regression estimator.

To learn more about the sampling distribution of the OLS estimator, we must make further assumptions about the distribution of the stochastic disturbance u_i. In classical statistics, the u_i were assumed to be independent draws from the same normal distribution. The modern approach to econometrics drops the normality assumption and simply assumes that the u_i satisfy the assumptions of a zero conditional mean and conditional homoskedasticity.

5. The treatment here is similar to that of Wooldridge (2025). See Stock and Watson (2019) for a derivation based on minimizing the squared-prediction errors.

The exact finite-sample distribution of the OLS estimator can be derived by using the normality assumption. In contrast, under these weaker assumptions, we must use large-sample theory to derive the sampling distribution of the OLS estimator. Basically, large-sample theory supposes that the sample size N becomes infinitely large. Because no real sample is infinitely large, these methods approximate the sampling distribution of the OLS estimator only in finite samples. With a few hundred observations or more, the large-sample approximation works well, so these methods work well with applied economic datasets.

Although large-sample theory is more abstract than finite-sample methods, it imposes weaker assumptions on the data-generating process. We will use large-sample theory to define "good" estimators and to evaluate the precision of the estimates produced from a given sample.

In large samples, consistency means that as N goes to ∞, the estimates will converge to their respective population parameters. Roughly speaking, if the probability that the estimator produces estimates arbitrarily close to the population values goes to 1 as the sample size increases to infinity, the estimator is said to be consistent.

The sampling distribution of an estimator describes the set of estimates produced when that estimator is applied to repeated samples from the underlying population. You can use the sampling distribution of an estimator to evaluate the precision of a given set of estimates and to statistically test whether the population parameters take on certain values.

Large-sample theory shows that the sampling distribution of the OLS estimator is approximately normal.[6] Specifically, when the u_i are conditionally homoskedastic with finite variance σ_u^2, the OLS estimator $\widehat{\boldsymbol{\beta}}$ has a large-sample normal distribution with mean β and variance $\sigma_u^2\mathbf{Q}^{-1}$, where $\mathbf{Q}^{-1}$ is the variance–covariance matrix of X in the population. The variance–covariance matrix of the estimator, $\sigma_u^2\mathbf{Q}^{-1}$, is also referred to as a VCE. Because it is unknown, we need a consistent estimator of the VCE. Although neither σ_u^2 nor $\mathbf{Q}^{-1}$ is actually known, we can use consistent estimators for them to construct a consistent estimator of $\sigma_u^2\mathbf{Q}^{-1}$. Given that s^2 consistently estimates σ_u^2 and $(1/N)(\mathbf{X}'\mathbf{X})$ consistently estimates $\mathbf{Q}$, $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is a VCE of the OLS estimator.[7]

4.2.3 Efficiency of the regression estimator

Under the assumption of conditionally homoskedastic errors, the Gauss–Markov theorem holds. Among linear, unbiased estimators of β, the OLS estimator has the smallest

6. More precisely, the distribution of the OLS estimator converges to a normal distribution. In this book, I will simply refer to the "approximate" or "large-sample" normal distribution. See Wooldridge (2025, chap. 5) for an introduction to large-sample theory.

7. At first glance, you might think that the expression for the VCE should be multiplied by $1/N$, but this assumption is incorrect. Because the OLS estimator is consistent, it is converging to the constant vector of population parameters at the rate $1/\sqrt{N}$, implying that the variance of the OLS estimator is going to 0 as the sample size gets larger. Large-sample theory compensates for this effect in how it standardizes the estimator. The loss of the $1/N$ term in the estimator of the VCE is a product of this standardization.

sampling variance, or the greatest precision.[8] In that sense, it is best, so OLS is "BLUE" (the best linear unbiased estimator) for the parameters of the regression model. If we consider only unbiased estimators that are linear in the parameters, we cannot find a more efficient estimator. The property of efficiency refers to the precision of the estimator. If estimator A has a smaller sampling variance than estimator B, then estimator A is said to be relatively efficient. The Gauss–Markov theorem states that OLS is relatively efficient versus all other linear, unbiased estimators of the model. However, this statement rests upon the hypotheses of an appropriately specified model and an error process with a zero conditional mean, as specified in (4.2). Note that although unbiasedness does not rely upon the assumption of conditional homoskedasticity in (4.3), it is required to achieve the BLUE result of the Gauss–Markov theorem.

4.2.4 Numerical identification of the regression estimates

As in (4.5) above, the solution to the regression problem involves a set of k moment conditions, or equations to be jointly solved for the k parameter estimates $\widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$. When will these k parameter estimates be uniquely determined, or numerically identified? We must have more sample observations than parameters to be estimated; in other words, N must be greater than k. That condition is not sufficient, though. For the simple "two-variable" regression model $y_i = \beta_1 + \beta_2 x_{i,2} + u_i$, $\mathrm{Var}(x_2)$ must be greater than 0. If there is no variation in x_2, the data do not provide sufficient information to determine estimates of β_1 and β_2.

In multiple regression with many regressors, $\mathbf{X}_{N\times k}$ must be a matrix of full column rank k, which implies two things. First, only one column of $\mathbf{X}$ can take on a constant value, so each other regressor must have a positive sample variance. Second, there are no exact linear dependencies among the columns of the matrix $\mathbf{X}$. The assumption that $\mathbf{X}$ is of full column rank is often stated as "$(\mathbf{X}'\mathbf{X})$ is of full rank" or "$(\mathbf{X}'\mathbf{X})$ is nonsingular (or invertible)". If the matrix of regressors $\mathbf{X}$ contains k linearly independent columns, the cross-product matrix $(\mathbf{X}'\mathbf{X})$ will have rank k, its inverse will exist, and the parameters $\beta_1, \ldots, \beta_k$ in (4.5) will be numerically identified.[9] If numerical identification fails, the sample does not contain enough information for us to use the regression estimator on the model as it is specified. That regression function may be valid as a description of the data-generating process, but the particular sample may lack the necessary information to generate a regressor matrix of full column rank. Then we must either respecify the model or acquire another sample that contains the information needed to uniquely determine the regression estimates.

8. For a formal presentation of the Gauss–Markov theorem, see any econometrics text, for example, Wooldridge (2025, 95–96). The OLS estimator is said to be unbiased because $E(\widehat{\boldsymbol{\beta}}) = \beta$.
9. When computing infinite precision, we must be concerned with numerical singularity and a computer program's ability to reliably invert a matrix regardless of whether it is analytically invertible. As we discuss in section 4.3.6, computationally near-linear dependencies among the columns of $\mathbf{X}$ should be avoided.

4.3 Interpreting regression estimates

This section illustrates using regression by an example from a prototype research project and discusses how Stata presents regression estimates. Then it discusses how to recover the information displayed in Stata's estimation results for further computations within your program and how to combine this information with other estimates to present them in a table. The last subsection considers problems of numerical identification, or collinearity, that could appear when you are estimating the regression equation.

4.3.1 Research project: A study of wage determination

As an illustration, I present regression estimates from a model fit to 2,246 individuals' wage data, in which the response variable is the wage in US dollars per hour in 1988. These data are extracted from the National Longitudinal Survey of Young Women, aged 14–24 in 1968.

`nlsw88.dta` contains individuals' attributes that we expect would influence the wage they earn. Our research question relates to the effect of each of these attributes on wages. We explore the variables available in these data with `describe` (see [D] **describe**). The `notes` (see [D] **notes**) command provides further information:

```
. webuse nlsw88
(NLSW, 1988 extract)

. describe

Contains data from https://www.stata-press.com/data/r19/nlsw88.dta
 Observations:         2,246                  NLSW, 1988 extract
    Variables:            17                  1 May 2220 22:52
                                              (_dta has notes)
-------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------
idcode          int     %8.0g                 NLS ID
age             byte    %8.0g                 Age in current year
race            byte    %8.0g      racelbl    Race
married         byte    %8.0g      marlbl     Married
never_married   byte    %16.0g     nev_mar    Never married
grade           byte    %8.0g                 Current grade completed
collgrad        byte    %16.0g     gradlbl    College graduate
south           byte    %9.0g      southlbl   Lives in the south
smsa            byte    %9.0g      smsalbl    Lives in SMSA
c_city          byte    %16.0g     ccitylbl   Lives in a central city
industry        byte    %23.0g     indlbl     Industry
occupation      byte    %22.0g     occlbl     Occupation
union           byte    %8.0g      unionlbl   Union worker
wage            float   %9.0g                 Hourly wage
hours           byte    %8.0g                 Usual hours worked
ttl_exp         float   %9.0g                 Total work experience (years)
tenure          float   %9.0g                 Job tenure (years)
-------------------------------------------------------------------------------
Sorted by: idcode
```

```
. notes
_dta:
  1.  1988 data, extracted from National Longitudinal Survey of Young Women who
      were ages 14-24 in 1968 (NLSW).
  2.  This dataset is the result of extraction and processing by various people
      at various times.
  3.  For more information on the NLS, see http://www.bls.gov/nls/.
```

We note that most of the variables are of `int` or `byte` datatype, with several expressed as `float` values, containing a decimal point. Several of the integer variables also have value labels, identifying the coding of variables such as `race`.

We present the descriptive statistics with `summarize`:

```
. summarize
```

Variable	Obs	Mean	Std. dev.	Min	Max
idcode	2,246	2612.654	1480.864	1	5159
age	2,246	39.15316	3.060002	34	46
race	2,246	1.282725	.4754413	1	3
married	2,246	.6420303	.4795099	0	1
never_marr~d	2,246	.1041852	.3055687	0	1
grade	2,244	13.09893	2.521246	0	18
collgrad	2,246	.2368655	.4252538	0	1
south	2,246	.4194123	.4935728	0	1
smsa	2,246	.7039181	.4566292	0	1
c_city	2,246	.2916296	.4546139	0	1
industry	2,232	8.189516	3.010875	1	12
occupation	2,237	4.642825	3.408897	1	13
union	1,878	.2454739	.4304825	0	1
wage	2,246	7.766949	5.755523	1.004952	40.74659
hours	2,242	37.21811	10.50914	1	80
ttl_exp	2,246	12.53498	4.610208	.1153846	28.88461
tenure	2,231	5.97785	5.510331	0	25.91667

The statistics reveal that many of the integer variables are indicators, or binary variables, while others take on multiple values. The `union` variable has 368 missing values. Because most statistical commands use casewise deletion (removing any observation for which any variable is coded as missing), the number of observations used in a model including `union` will be limited. All the variables take on sensible values, with the possible exception of `grade`, which should not be 0. To investigate, we use `tabulate` to examine its distribution:

```
. tabulate grade
```

Current grade completed	Freq.	Percent	Cum.
0	2	0.09	0.09
4	3	0.13	0.22
5	1	0.04	0.27
6	14	0.62	0.89
7	19	0.85	1.74
8	33	1.47	3.21
9	55	2.45	5.66
10	84	3.74	9.40
11	123	5.48	14.88
12	943	42.02	56.91
13	176	7.84	64.75
14	187	8.33	73.08
15	92	4.10	77.18
16	252	11.23	88.41
17	106	4.72	93.14
18	154	6.86	100.00
Total	2,244	100.00	

We conjecture that it is implausible that respondents received less than a primary education. Therefore, we change values that are less than 6 to missing:

```
. replace grade = . if grade<6
(6 real changes made, 6 to missing)
```

We can also examine more detailed summary statistics for our outcome variable, `wage`:

```
. summarize wage, detail
```

Hourly wage

Percentiles		Smallest		
1%	1.930993	1.004952		
5%	2.801002	1.032247		
10%	3.220612	1.151368	Obs	2,246
25%	4.259257	1.344605	Sum of wgt.	2,246
50%	6.27227		Mean	7.766949
		Largest	Std. dev.	5.755523
75%	9.597424	40.19808		
90%	12.77777	40.19808	Variance	33.12604
95%	16.52979	40.19808	Skewness	3.096199
99%	38.70926	40.74659	Kurtosis	15.85446

The variable is heavily skewed with a long right tail, as expected. This can be viewed using the `kdensity` (see [R] **kdensity**) command, as shown in figure 4.2:

```
. kdensity wage
```

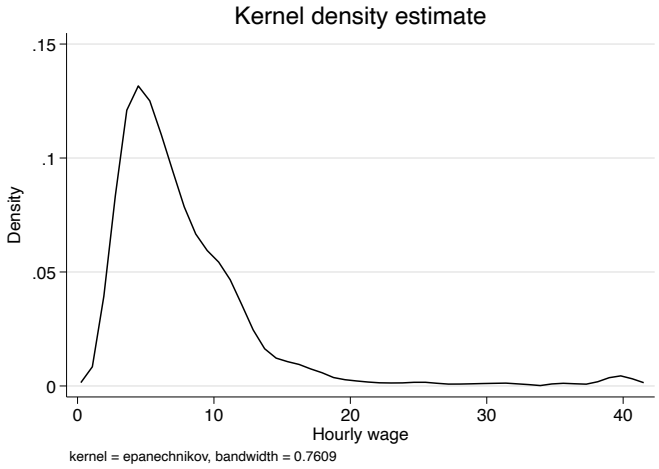


Figure 4.2. Density estimate of wage

Because analysis of wages often makes use of a log-linear model, we transform the wage with logarithms and reevaluate the density plot in figure 4.3:

```
. generate lwage = log(wage)
. label variable lwage "log(wage)"
. kdensity lwage, normal legend(position(6) rows(1))
```

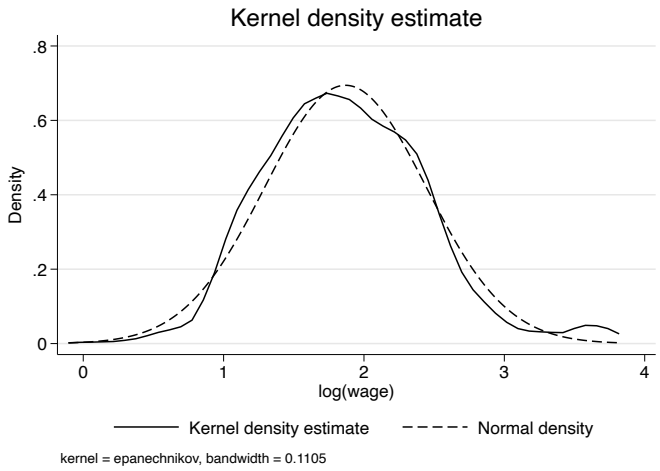


Figure 4.3. Density estimate of log(wage)

Using the `normal` option of `kdensity`, we see that although `wage` is highly skewed, the distribution of `lwage` is much closer to that of a normally distributed random variable. Although we can estimate a log-linear regression, the generalized linear model (see [R] **glm**) is an alternative estimator that may be preferable. We will discuss this model in section 5.3.

We may also want to explore the frequency distributions of some of the integer variables, such as `race` and `age`, using the `tabulate twoway` and `table` (see [R] **tabulate twoway** and [R] **table**) commands:

```
. tabulate age race
```

Age in current year	Race White	Black	Other	Total
34	37	15	1	53
35	183	74	3	260
36	184	71	2	257
37	146	75	4	225
38	158	59	2	219
39	165	67	2	234
40	167	41	0	208
41	162	56	4	222
42	121	37	2	160
43	129	33	3	165
44	117	43	3	163
45	66	12	0	78
46	2	0	0	2
Total	1,637	583	26	2,246

```
. table (age) (race), stat(percent,across(race)) nototal
```

	Race White	Black	Other
Age in current year			
34	69.81	28.30	1.89
35	70.38	28.46	1.15
36	71.60	27.63	0.78
37	64.89	33.33	1.78
38	72.15	26.94	0.91
39	70.51	28.63	0.85
40	80.29	19.71	
41	72.97	25.23	1.80
42	75.62	23.12	1.25
43	78.18	20.00	1.82
44	71.78	26.38	1.84
45	84.62	15.38	
46	100.00		

We see that Black respondents have a smaller share of the sample for higher age groups.

The `table` command can also be used to produce three-way tables:

```
. table (married collgrad) (race), nototal
```

	Race		
	White	Black	Other
Married			
Single			
College graduate			
Not college grad	355	256	5
College grad	132	53	3
Married			
College graduate			
Not college grad	862	224	12
College grad	288	50	6

We can also generate tables of summary statistics with `table`:

```
. table (result) (race), stat(mean wage) stat(sd wage) stat(p50 wage)
> nformat(%5.2f)
```

	Race			
	White	Black	Other	Total
Mean	8.08	6.84	8.55	7.77
Standard deviation	5.96	5.08	5.21	5.76
50th percentile	6.55	5.43	7.56	6.27

We see that White respondents have a considerably higher mean wage than Black respondents. This also holds for the median (`p50`) wage.

Now that we have familiarized ourselves with the data, we fit a multiple regression of `wage` on three continuous variables: `ttl_exp`, `tenure`, and `grade`. We use the `i.` factor-variable prefix on a list of integer variables to be included: `race`, `south`, `smsa`, and `union`. The `regress` (see [R] **regress**) command, like other Stata estimation commands, requires us to specify the response variable followed by a list of the explanatory variables:[10]

```
. regress wage ttl_exp tenure grade i.(race south smsa union)
```

Source	SS	df	MS	Number of obs	=	1,861
				F(8, 1852)	=	112.00
Model	10538.705	8	1317.33812	Prob > F	=	0.0000
Residual	21783.4718	1,852	11.7621338	R-squared	=	0.3261
				Adj R-squared	=	0.3231
Total	32322.1768	1,860	17.3775144	Root MSE	=	3.4296

wage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ttl_exp	.232707	.0215614	10.79	0.000	.1904198	.2749942
tenure	.0522526	.0175946	2.97	0.003	.0177453	.08676
grade	.5527474	.0334101	16.54	0.000	.487222	.6182728
race						
Black	-.6309841	.1929753	-3.27	0.001	-1.009456	-.2525121
Other	.5670115	.7232603	0.78	0.433	-.8514796	1.985503
south						
South	-.6525186	.1726751	-3.78	0.000	-.9911769	-.3138603
smsa						
SMSA	1.51067	.1805822	8.37	0.000	1.156504	1.864836
union						
Union	.7640577	.1903099	4.01	0.000	.3908131	1.137302
_cons	-3.850026	.4807968	-8.01	0.000	-4.792987	-2.907066

The header of the regression output describes the overall model estimates, whereas the table presents the point estimates, their precision, and their interval estimates.

4.3.2 The ANOVA table: ANOVA F and R^2

The regression output for this model includes the analysis of variance (ANOVA) table in the upper left, where the two sources of variation are displayed as `Model` and `Residual`. The `SS` column shows the sums of squares, with the `Residual SS` corresponding to $\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$ and the `Total SS` to $\widetilde{\mathbf{y}}'\widetilde{\mathbf{y}}$ in (4.7) below. The `df` column reports the degrees of freedom associated with each sum of squares. The degrees of freedom for the total `SS` is $(N-1)$ because the total `SS` have been computed by using one sample statistic, $\overline{y}$. The degrees of freedom for the model is $(k-1)$, equal to the number of slopes (or explanatory variables), or one fewer than the number of estimated coefficients due to the constant term. `Model SS` refers to the ability of the regressors to jointly explain a

10. Without any explanatory variables, **regress** estimates the sample mean of the response variable.

fraction of the variation of y about its mean (`Total SS`). The residual degrees of freedom is $(N-k)$, indicating that $(N-k)$ residuals may be freely determined and still satisfy the constraint from the first normal equation of least squares that the regression surface passes through the multivariate point of means $(\overline{y}, \overline{x}_2, \ldots, \overline{x}_k)$:

$$\overline{y} = \widehat{\beta}_1 + \widehat{\beta}_2\overline{x}_2 + \widehat{\beta}_3\overline{x}_3 + \cdots + \widehat{\beta}_k\overline{x}_k$$

In the presence of the constant term $\widehat{\beta}_1$, the first normal equation implies that $\overline{\widehat{u}} = \overline{y} - \Sigma_i \overline{\mathbf{x}}_i \widehat{\boldsymbol{\beta}}_i$ must be identically 0.[11] This is not an assumption but rather an algebraic implication of the least-squares technique, which guarantees that the sum and mean of least-squares residuals will be close to 0.[12]

The `MS` column reports the mean squares due to regression and error, or the `SS` divided by the `df`. The ratio of the `Model MS` to `Residual MS` is reported as the ANOVA F statistic, with numerator and denominator degrees of freedom equal to the respective `df` values. This ANOVA F statistic is a test of the null hypothesis[13] that the slope coefficients in the model are jointly 0; that is, the null model of $y_i = \mu + u_i$ is as successful in describing y as is the regression alternative. `Prob > F` is the tail probability or p-value of the F statistic. Here we can reject the null hypothesis at any conventional level of significance. Also, the `Root MSE` for the regression of 3.43, which is in the units of the response variable y, can be compared with the mean of that variable, 7.60. We can use the `estat summarize` postestimation command to evaluate the values of the observations included in the model because the inclusion of `union` has reduced the size of the estimation sample to 1,861 observations.

11. Recall that the first column of $\mathbf{X} = \iota$, an N-element unit vector.
12. Because computers use finite arithmetic, the sum will differ from 0. A well-written computer program should result in a difference similar to machine precision. For this regression, Stata reports a mean residual of -1.4×10^{-15}, comparable with the `epsdouble()` value of 2.2×10^{-16}, which is the smallest number distinguishable by Stata.
13. I discuss hypothesis testing in detail in section 4.5.

```
. estat summarize
Estimation sample regress                       Number of obs =      1,861
```

Variable	Mean	Std. dev.	Min	Max
wage	7.600456	4.168635	1.344605	39.23074
ttl_exp	12.85004	4.596992	.1153846	28.88461
tenure	6.571557	5.638432	0	25.91667
grade	13.19882	2.494557	6	18
race				
Black	.2649113	.4414046	0	1
Other	.0123589	.1105114	0	1
south				
South	.424503	.4944002	0	1
smsa				
SMSA	.7082214	.4547031	0	1
union				
Union	.2466416	.4311721	0	1

The upper-right section of the `regress` output contains several goodness-of-fit statistics, which measure the degree to which a fitted model can explain the variation of the response variable y. All else equal, we should prefer a simpler model with a better fit to the data. The mechanics of regression imply that a model with many regressors can explain y arbitrarily well. Given the least-squares residuals, the most common measure of goodness of fit, regression R^2, may be calculated (given a constant term in the regression function) as

$$R^2 = 1 - \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{\widetilde{\mathbf{y}}'\widetilde{\mathbf{y}}} \tag{4.7}$$

where $\widetilde{\mathbf{y}} = y - \overline{y}$: the regressand with its sample mean removed. This calculation emphasizes that the object of regression is not to explain $\mathbf{y}'\mathbf{y}$, which is the raw sum of squares of the response variable y. This would merely explain why $E(y) \neq 0$—not an interesting question. Rather, we want to explain the variation in the response variable around its sample mean.

With a constant term in the model, the least-squares approach seeks to explain the largest possible fraction of the sample variation of y about its mean and not the associated variance. The null model with which (4.1) is contrasted is $y = \mu + u_i$, where μ is the population mean of y. When we fit a regression, we want to determine whether the information in the regressors $\mathbf{x}$ is useful. Is the conditional expectation $E(y|\mathbf{x})$ more informative than the unconditional expectation $E(y) = \mu$? The null model above has an $R^2 = 0$, whereas virtually any set of regressors will explain some fraction of the variation of y around $\overline{y}$, the sample estimate of μ. R^2 is that fraction in the unit interval, the proportion of the variation in y about $\overline{y}$ explained by $\mathbf{x}$.

(Pages omitted)