An Introduction to Modern Econometrics Using Stata

CHRISTOPHER F. BAUM
Department of Economics
Boston College

A Stata Press Publication
StataCorp LP
College Station, Texas

Stata Press, 4905 Lakeway Drive, College Station, Texas 77845

Typeset in LaTeX 2ε
Printed in the United States of America
10 9 8 7 6 5 4 3 2 1

ISBN-10: 1-59718-013-0
ISBN-13: 978-1-59718-013-9

(Pages omitted)

Contents

(Pages omitted)

Preface

This book is a concise guide for applied researchers in economics and finance to learn basic econometrics and use Stata with 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, or OLS) model and its algebraic representation, equivalent to the level of an undergraduate statistics/econometrics course sequence.[1] 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 should consult the appropriate *Getting Started with Stata* manual if review is needed. Meanwhile, readers already comfortable interacting with Stata should feel free to skip to chapter 4, where the discussion of econometrics begins in earnest.

In any research project, a great deal of the effort is involved with the preparation of the data specified as part of an econometric model. While the primary focus of the book is placed upon applied econometric practice, we must consider the considerable challenges that many researchers face in moving from their original data sources to the form needed in an econometric model—or even that 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, you should feel free to skim this material, perhaps returning to it to refresh your understanding of Stata usage. Likewise, Chapter 3 is devoted to a discussion of the organization of economic and financial data, and the Stata commands needed to reorganize data among the several forms of organization (cross section, time series, pooled, panel/longitudinal, etc.) If you are eager to begin with the econometrics of linear regression, skim this chapter, noting its content for future reference.

Chapter 4 begins the econometric content of the book and presents the most widely used tool for econometric analysis: the multiple linear regression model applied to continuous 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 may be violated in the presence of specification error. The chapter also

[1] Two excellent texts at this level are Wooldridge (2006) and Stock and Watson (2006).

discusses statistical and graphical techniques for detecting specification error. Chapter 6 discusses other assumptions that may be violated, such as the assumption of independent and identically distributed (i.i.d.) errors, and presents the generalized linear regression model. It also explains how to diagnose and correct the two most important departures from i.i.d., heteroskedasticity and serial correlation.

Chapter 7 discusses using indicator variables or dummy variables in the 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, OLS 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, or a constrained response such as the amount spent, which 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 appendices 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.

(Pages omitted)

4 Linear regression

This chapter presents 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. lay out 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. On the basis of these findings, we may have to return to one of the earlier stages to reevaluate the dataset and its specification and functional form.

Section 2 reviews the basic regression analysis theory on which regression point and interval estimates are based. Section 3 introduces a prototype economic research project studying the determinants of communities' single-family housing prices and discusses the various components of Stata's results from fitting a regression model of housing prices. Section 4 discusses how to transform Stata's estimation results into publication-quality tables. Section 5 discusses hypothesis testing and estimation subject to constraints on the parameters. Section 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

$$E\left[y \mid x_1, x_2, \ldots, x_k\right] = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Given values for the βs, which are fixed parameters, the linear regression model 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 single-family home prices in Boston-area communities, conditional on the student–teacher ratios, is given by

$$E\left[\texttt{price} \mid \texttt{stratio}\right] = \beta_1 + \beta_2\ \texttt{stratio}$$

where `price` is the mean value of single-family home prices and `stratio` is the student–teacher ratio. This relationship reflects the hypothesis that the quality of communities' school systems is capitalized into housing prices. Here the population is the set of communities in the Commonwealth of Massachusetts. Each town or city in Massachusetts is generally responsible for its own school system.

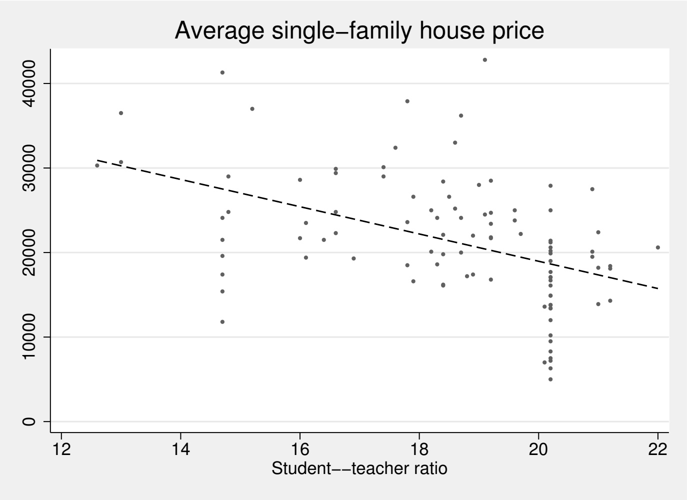

Figure 4.1: Conditional mean of single-family house price

Figure 4.1 shows average single-family housing prices for 100 Boston-area communities, along with the linear fit of housing prices to student–teacher ratios. The conditional mean of `price` for each value of `stratio` is shown by the appropriate point on the line. As theory predicts, the mean house price conditional on the student–teacher ratio is inversely related to that ratio. Communities with more crowded schools are considered less desirable. Of course, this relationship between house price and the student–teacher ratio must be considered ceteris paribus: all other factors that might affect the price of the house are held constant when we evaluate the effect of a measure of community schools' quality on the house price.

In 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 good 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 model 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 must be much larger than k.

We can write the linear regression model 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)

[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 $(k+1)$.

y. We usually assume that the model contains a constant term, so x_1 is understood to equal one for each observation.

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

$$E\left[u \mid \mathbf{x}\right] = 0 \tag{4.2}$$

i.e., 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 the regression model lets us consider both nonstochastic and stochastic regressors in $\mathbf{X}$ without distinction, as long as they satisfy the assumption of (4.2).[3]

4.2.1 Regression as a method-of-moments estimator

We may 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]

$$\begin{aligned} E[\mathbf{x}'u] &= \mathbf{0} \\ E[\mathbf{x}'(y - \mathbf{x}\boldsymbol{\beta})] &= \mathbf{0} \end{aligned} \tag{4.3}$$

Substituting calculated moments from our sample into the expression and replacing the unknown coefficients $\boldsymbol{\beta}$ with estimated values $\widehat{\boldsymbol{\beta}}$ in (4.3) yields the *ordinary least squares* (OLS) estimator

$$\begin{aligned} \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} &= \mathbf{0} \\ \widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned} \tag{4.4}$$

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

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

[2] $\mathbf{x}$ is a vector of random variables and u is scalar random variable. In (4.1), $\mathbf{X}$ is a matrix of realizations of the random vector $\mathbf{x}$, $\mathbf{u}$ 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 encountered.

[4] The assumption of zero-conditional mean is stronger than that of a zero covariance, because covariance considers only *linear* relationships between the random variables.

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—may be estimated as a function of the regression residuals $\widehat{u}_i$

$$s^2 = \frac{\sum_{i=1}^{N} \widehat{u}_i^2}{N-k} = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{N-k} \tag{4.5}$$

where $(N-k)$ are 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 (4.4), 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 are independent draws from an identical distribution (i.i.d.).[6]

Using the normality assumption, we were able to derive the exact finite-sample distribution of the OLS estimator. In contrast, under the i.i.d. assumption, 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. Since no real sample is infinitely large, these methods only approximate the sampling distribution of the OLS estimator 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.

[5] The treatment here is similar to that of Wooldridge (2006). See Stock and Watson (2006) and appendix 4.A for a derivation based on minimizing the squared-prediction errors.

[6] Both frameworks also assume that the (constant) variance of the u process is finite. Formally, i.i.d. stands for independently and identically distributed.

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 one 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.[7] Specifically, when the u_i are i.i.d. 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 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 of 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.[8]

4.2.3 Efficiency of the regression estimator

Under the assumption of i.i.d. errors, the Gauss–Markov theorem holds. Among linear, unbiased estimators, the OLS estimator has the smallest sampling variance, or the greatest precision.[9] In that sense, it is *best*, so that "ordinary least squares 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, estimator A is said to be *relatively efficient*. The Gauss–Markov theorem states that OLS is relatively efficient

[7] More precisely, the distribution of the OLS estimator converges to a normal distribution. Although appendix B provides some details, in the text I will simply refer to the "approximate" or "large-sample" normal distribution. See Wooldridge (2006) for an introduction to large-sample theory.

[8] At first glance, you might think that the expression for the VCE should be multiplied by $1/N$, but this assumption is incorrect. As discussed in appendix B, 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 zero 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.

[9] For a formal presentation of the Gauss–Markov theorem, see any econometrics text, e.g., Wooldridge (2006, 108–109). The OLS estimator is said to be "unbiased" because $E[\widehat{\boldsymbol{\beta}}] = \beta$.

versus all other linear, unbiased estimators of the parameterization model. However, this statement rests upon the hypotheses of an appropriately specified model and an i.i.d. disturbance process with a zero-conditional mean, as specified in (4.2).

4.2.4 Numerical identification of the regression estimates

As in (4.4) 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, or $N > k$. That condition is not sufficient, though. For the simple "two-variable" regression model $y_i = \beta_1 + \beta_2 x_{i,2} + u_i$, $\text{Var}[x_2]$ 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 of the other regressors 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 (4.4) will be numerically identified.[10] 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 model 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.

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. We then discuss 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 may appear when you are estimating the regression equation.

[10] 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 4.3.7, computationally *near-linear dependencies* among the columns of $\mathbf{X}$ should be avoided.

4.3.1 Research project: A study of single-family housing prices

As an illustration, we present regression estimates from a model fitted to 506 Boston-area communities' housing price data, in which the response variable is the logarithm of the median price of a single-family home in each community. The dataset (`hprice2a`) contains an attribute of each community's housing stock that we would expect to influence price: `rooms`, the average number of rooms per house. Our research question relates to the influences on price exerted by several external factors. These factors, measured at the community level, include a measure of air pollution (`lnox`, the log of nitrous oxide in parts per 100m), the distance from the community to employment centers (`ldist`, the log of the weighted distance to five employment centers), and the average student–teacher ratio in local schools (`stratio`). From economic theory, we would expect the average number of rooms to increase the price, ceteris paribus. Each of the external factors is expected to decrease the median housing price in the community. More polluted communities, those less conveniently situated to available jobs, and those with poorly staffed schools should all have less expensive housing, given the forces of supply and demand.

We present the descriptive statistics with `summarize` and then fit a regression equation.

```
. use http://www.stata-press.com/data/imeus/hprice2a, clear
(Housing price data for Boston-area communities)

. summarize price lprice lnox ldist stratio, sep(0)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |       506    22511.51    9208.856       5000      50001
      lprice |       506    9.941057     .409255   8.517193    10.8198
        lnox |       506    1.693091    .2014102   1.348073   2.164472
       ldist |       506    1.188233     .539501   .1222176   2.495682
     stratio |       506    18.45929     2.16582       12.6         22
```

The `regress` command, like other Stata estimation commands, requires us to specify the response variable followed by a *varlist* of the explanatory variables.

```
. regress lprice lnox ldist rooms stratio

      Source |       SS       df       MS              Number of obs =     506
-------------+------------------------------           F(  4,   501) =  175.86
       Model |  49.3987735     4  12.3496934           Prob > F      =  0.0000
    Residual |  35.1834974   501  .070226542           R-squared     =  0.5840
-------------+------------------------------           Adj R-squared =  0.5807
       Total |  84.5822709   505  .167489645           Root MSE      =    .265

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lnox |    -.95354   .1167418    -8.17   0.000    -1.182904   -.7241762
       ldist |  -.1343401   .0431032    -3.12   0.002    -.2190255   -.0496548
       rooms |   .2545271   .0185303    13.74   0.000     .2181203    .2909338
     stratio |  -.0524512   .0058971    -8.89   0.000    -.0640373   -.0408651
       _cons |   11.08387   .3181115    34.84   0.000     10.45887    11.70886
------------------------------------------------------------------------------
```

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-squared

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` are the sums of squares, with the `Residual SS` corresponding to $\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$ and the total `Total SS` to $\widetilde{\mathbf{y}}'\widetilde{\mathbf{y}}$ in (4.6) below. The next column of the table reports the `df`: the degrees of freedom associated with each sum of squares. The degrees of freedom for total `SS` are $(N-1)$ since the total `SS` have been computed by using one sample statistic, $\overline{y}$. The degrees of freedom for the model are $(k-1)$, equal to the number of slopes (or explanatory variables), or one fewer than the number of estimated coefficients due to the constant term. The model `SS` refer to the ability of the four regressors to jointly explain a fraction of the variation of y about its mean (the total `SS`). The residual degrees of freedom are $(N-k)$, 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 zero.[11] This is not an assumption but is an algebraic implication of the least-squares technique, which guarantees that the sum of least-squares residuals (and their mean) will be very close to zero.[12]

The last column of the ANOVA table reports the `MS`, 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 zero: that is, the null model of $y_i = \mu + u_i$ is as successful in describing y as the regression alternative. The `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 0.265, which is in the units of the response variable y, is small relative to the mean of that variable, 9.94.

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 model with a better fit to the data. For the sake of parsimony, we also prefer a simpler model. The mechanics of

[11] Recall that the first column of $\mathbf{X} = \iota$, an N-element unit vector.

[12] Since computers use finite arithmetic, the sum will differ from zero. 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 to 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.

regression imply that a model with a great 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

$$R^2 = 1 - \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{\widetilde{\mathbf{y}}'\,\widetilde{\mathbf{y}}} \tag{4.6}$$

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 $\mathbf{y}'\mathbf{y}$, the raw sum of squares of the response variable y, which would merely explain why $E[y] \neq 0$—not an interesting question. Rather, the object is to explain the variations in the response variable.

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. In estimating 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 $\mathbf{x}$.

4.3.3 Adjusted R-squared

What about the `Adj R-squared`? The algebra of least squares dictates that adding a $(k+1)$st column to $\mathbf{X}$ will result in a regression estimate with $R^2_{k+1} \geq R^2_k$. R^2 cannot fall with the addition of $\mathbf{x}_{k+1}$ to the regression equation, as long as the observations on the marginal regressor are linearly independent of the previous k columns from a numerical standpoint.[14] Indeed, we know that R^2_N (that is, R^2 calculated from a regression in which there are N linearly independent columns of $\mathbf{X}$ and N observations in the sample) must equal 1.0. As we add regressors to $\mathbf{x}$, R^2 cannot fall and is likely to rise, even when the marginal regressor is irrelevant econometrically.

What if we have a competing model that cannot be expressed as nested within this model, and this model does not nest within the competing model? A nonstatistical approach to this problem, especially where the two models differ widely in their numbers of regressors (or `Model df`), is to consider their $\overline{R}^2$ values, the statistic Stata labels as `Adj R-squared`.[15] The $\overline{R}^2$ considers the explained *variance* of y, rather than the explained *variation*, as does ordinary R^2. That is, rather than merely considering $\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$, the residual sum of squares, $\overline{R}^2$ takes into account the degrees of freedom lost in fitting

[14] In this sense, the limitations of finite arithmetic using the binary number system intrude: since 0.100 cannot be exactly expressed in a finite number of digits in the binary system, even a column that should be perfectly collinear with the columns of $\mathbf{X}_k$ may not be so computationally. The researcher should know her data and recognize when a candidate regressor cannot logically add information to an existing regressor matrix, whether or not the resulting regressor matrix is judged to possess full column rank by Stata.

[15] A formal statistical approach to the nonnested models problem is presented below in section 4.5.5.

the model and scales $\widehat{\mathbf{u}}'\widehat{\mathbf{u}}$ by $(N-k)$ rather than N.[16] $\overline{R}^2$ can be expressed as a corrected version of R^2 in which the degrees-of-freedom adjustments are made, penalizing a model with more regressors for its loss of parsimony:

$$\overline{R}^2 = 1 - \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}/(N-k)}{\widetilde{\mathbf{y}}'\,\widetilde{\mathbf{y}}/(N-1)} = 1 - (1-R^2)\frac{N-1}{N-k}$$

If an irrelevant regressor is added to a model, R^2 cannot fall and will probably rise, but $\overline{R}^2$ will rise if the benefit of that regressor (reduced variance of the residuals) exceeds the cost of including it in the model: 1 degree of freedom.[17] Therefore, $\overline{R}^2$ can fall when a more elaborate model is considered, and indeed it is not bounded by zero. Algebraically, $\overline{R}^2$ must be less than R^2 since $(N-1)/(N-k) > 1$ for any $\mathbf{X}$ matrix and cannot be interpreted as the "proportion of variation of y", as can R^2 in the presence of a constant term. Nevertheless, you can use $\overline{R}^2$ to informally compare models with the same response variable but differing specifications. You can also compare the equations' s^2 values (labeled `Root MSE` in Stata's output) in units of the dependent variable to judge nonnested specifications.

Two other measures commonly used to compare competing regression models are the Akaike information criterion (AIC; Akaike [1974]) and Bayesian information criterion (BIC; often referred to as the Schwarz criterion: Schwarz [1978]). These measures also account for both the goodness of fit of the model and its parsimony. Each measure penalizes a larger model for using additional degrees of freedom while rewarding improvements in goodness of fit. The BIC places a higher penalty on using degrees of freedom. You can calculate the AIC and BIC after a regression model with the `estat ic` command. `estat ic` will display the log likelihood of the null model (that with only a constant term), the log likelihood of the fitted model, the model degrees of freedom, and the AIC and BIC values. For the regression above, we would type

```
. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    506   -265.4135   -43.49514      5     96.99028     118.123
-----------------------------------------------------------------------------
```

Least-squares regression can also be considered a maximum likelihood estimator of the vector $\boldsymbol{\beta}$ and ancillary parameter σ_u^2.[18] The degree to which our fitted model improves upon the null model in explaining the variation of the response variable is measured by the (algebraically) larger magnitude of `ll(model)` than that of `ll(null)`.[19]

[16] For comparison you may write (4.6), dividing both numerator and denominator by N.

[17] This is not a statistical judgment, as $\overline{R}^2_{k+1}$ can exceed $\overline{R}^2_k$ if the t statistic on the added regressor exceeds 1.0 in absolute value.

[18] The maximum likelihood estimator requires the normality assumption. See Johnston and DiNardo (1997).

[19] A *likelihood-ratio test* formally compares these two magnitudes under the null hypothesis that the null model is adequate. I discuss likelihood-ratio tests in chapter 10.

4.3.4 The coefficient estimates and beta coefficients

Below the ANOVA table and summary statistics, Stata reports the $\widehat{\boldsymbol{\beta}}$ coefficient estimates, along with their estimated standard errors, t statistics, and the associated p-values labeled `P>|t|`: that is, the tail probability for a two-tailed test on the hypothesis H_0: $\widehat{\beta}_j = 0$.[20] The last two columns display an estimated confidence interval, with limits defined by the current setting of `level`. You can use the `level()` option on `regress` (or other estimation commands) to specify a particular level. After performing the estimation (e.g., with the default 95% level), you can redisplay the regression results with, for instance, `regress, level(90)`. You can change the default `level` (see [R] **level**) for the session or permanently with `set level` # $\big[$ `, permanently` $\big]$.

Economic researchers often express regressors or response variables in logarithms.[21] A model in which the response variable is the log of the original series and the regressors are in levels is termed a *log-linear* (or *single-log*) model. The rough approximation that $\log(1 + x) \simeq x$ for reasonably small x is used to interpret the regression coefficients. These coefficients are also the *semielasticities* of y with respect to x, measuring the response of y in percentage terms to a unit change in x. When logarithms are used for both the response variable and regressors, we have the *double-log* model. In this model, the coefficients are themselves *elasticities* of y with respect to each x. The most celebrated example of a double-log model is the Cobb–Douglas production function, $q = al^{\alpha}k^{\beta}e^{\epsilon}$, which we can estimate by linear regression by taking logs of q, l, and k.

In other social science disciplines, linear regression results are often reported as estimated *beta coefficients*. This terminology is somewhat confusing for economists, given their common practice of writing the regression model in terms of βs. The beta coefficient is defined as $\partial y^*/\partial x_j^*$, where the starred quantities are z-transformed or standardized variables; for instance, $y^* = (y_i - \overline{y})/s_y$, where $\overline{y}$ is the sample mean and s_y is the sample standard deviation of the response variable. Thus the beta coefficient for the jth regressor tells us how many standard deviations y would change given a 1–standard deviation change in x_j. This measure is useful in disciplines where many empirical quantities are indices lacking a natural scale. We can then rank regressors by the magnitudes of their beta coefficients because the absolute magnitude of the beta coefficient for x_j indicates the strength of the effect of that variable. For the regression model above, we can merely redisplay the regression by using the `beta` option:

[20] We discuss hypothesis testing in detail in section 4.5.

[21] Economists use natural logs exclusively; references to `log` should be taken as the natural log, or `ln`.

```
. regress, beta
      Source |       SS       df       MS              Number of obs =     506
-------------+------------------------------           F(  4,   501) =  175.86
       Model |  49.3987735     4  12.3496934           Prob > F      =  0.0000
    Residual |  35.1834974   501  .070226542           R-squared     =  0.5840
-------------+------------------------------           Adj R-squared =  0.5807
       Total |  84.5822709   505  .167489645           Root MSE      =    .265

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
        lnox |    -.95354   .1167418    -8.17   0.000                -.4692738
       ldist |  -.1343401   .0431032    -3.12   0.002                -.1770941
       rooms |   .2545271   .0185303    13.74   0.000                 .4369626
     stratio |  -.0524512   .0058971    -8.89   0.000                -.2775771
       _cons |   11.08387   .3181115    34.84   0.000                        .
------------------------------------------------------------------------------
```

The output indicates that `lnox` has the largest beta coefficient, in absolute terms, followed by `rooms`. In economic and financial applications, where most regressors have a natural scale, it is more common to compute marginal effects such as elasticities or semielasticities (see section 4.7).

4.3.5 Regression without a constant term

With Stata, you can estimate a regression equation without a constant term by using the `noconstant` option, but I do not recommend doing so. Such a model makes little sense if the mean of the response variable is nonzero and all regressors' coefficients are insignificant.[22] Estimating a constant term in a model that does not have one causes a small loss in the efficiency of the parameter estimates. In contrast, incorrectly omitting a constant term produces inconsistent estimates. The tradeoff should be clear: include a constant term, and let the data indicate whether its estimate can be distinguished from zero.

What if we want to estimate a homogeneous relationship between y and the regressors $\mathbf{x}$, where economic theory posits $y \propto \mathbf{x}$? We can test the hypothesis of proportionality by estimating the relationship with a constant term and testing $H_0: \beta_1 = 0$. If the data reject that hypothesis, we should not fit the model with the constant term removed. Many of the common attributes of a linear regression are altered in a model that truly lacks a constant term. For instance, the least-squares residuals are not constrained to have zero sum or mean, and R^2 measured conventionally will be negative when the null model $y_i = \mu + u_i$ is not only preferable but strictly dominates the model $y_i = \beta_2 x_{i,2} + u_i$. Therefore, unless we have a good reason to fit a model without a constant term, we should retain the constant. An estimated $\widehat{\beta}_1$ not significantly different from zero does not harm the model, and it renders the model's summary statistics comparable to those of other models of the response variable y.

[22] If we provide the equivalent of a constant term by including a set of regressors that add up to a constant value for each observation, we should specify the `hascons` option as well as `noconstant`. Using the `hascons` option will alter the `Model SS` and `Total SS`, affecting the ANOVA F and R^2 measures; it does not affect the `Root MSE` or the t statistics for individual coefficients.