I absorbed the basic theory behind the Meltzer and Richard project, though I didn't find it all that plausible, since voter turnout is low for the poor. I also remember arguing with my bosses over whether government expenditure on education should be classified as a public good (something that benefits everyone in society as well as those directly affected) or a private good publicly supplied, and therefore a form of redistribution like welfare. You might say this project marked the beginning of my interest in the social returns to education, a topic I went back to with more enthusiasm and understanding in Acemoglu and Angrist (2000).

Today, I understand the Meltzer and Richard study as an attempt to use regression to uncover and quantify an interesting causal relation. At the time, however, I was purely a regression mechanic. Sometimes I found the RA work depressing. Days would go by when I didn't talk to anybody but my bosses and the occasional Carnegie-Mellon Ph.D. student, most of whom spoke little English anyway. The best part of the job was lunch with Allan Meltzer, a distinguished scholar and a patient and good-natured supervisor, who was happy to chat while we ate the contents of our brown bags (this did not take long, as Allan ate little and I ate fast). Once I asked Allan whether he found it satisfying to spend his days perusing regression output, which then came on reams of double-wide green-bar paper. Meltzer laughed and said there was nothing he would rather be doing.

Now we too spend our days happily perusing regression output, in the manner of our teachers and advisers in college and graduate school. This chapter explains why.

3.1 Regression Fundamentals

The end of the previous chapter introduced regression models as a computational device for the estimation of treatment-control differences in an experiment, with and without covariates. Because the regressor of interest in the class size study discussed in section 2.3 was randomly assigned, the resulting estimates have a causal interpretation. In most studies,

however, regression is used with observational data. Without the benefit of random assignment, regression estimates may or may not have a causal interpretation. We return to the central question of what makes a regression causal later in this chapter.

Setting aside the relatively abstract causality problem for the moment, we start with the mechanical properties of regression estimates. These are universal features of the population regression vector and its sample analog that have nothing to do with a researcher's interpretation of his output. These properties include the intimate connection between the population regression function and the conditional expectation function and the sampling distribution of regression estimates.

3.1.1 Economic Relationships and the Conditional Expectation Function

Empirical economic research in our field of labor economics is typically concerned with the statistical analysis of individual economic circumstances, and especially differences between people that might account for differences in their economic fortunes. Differences in economic fortune are notoriously hard to explain; they are, in a word, random. As applied econometricians, however, we believe we can summarize and interpret randomness in a useful way. An example of "systematic randomness" mentioned in the introduction is the connection between education and earnings. On average, people with more schooling earn more than people with less schooling. The connection between schooling and earnings has considerable predictive power, in spite of the enormous variation in individual circumstances that sometimes clouds this fact. Of course, the fact that more educated people tend to earn more than less educated people does not mean that schooling *causes* earnings to increase. The question of whether the earnings-schooling relationship is causal is of enormous importance, and we come back to it many times. Even without resolving the difficult question of causality, however, it's clear that education predicts earnings in a narrow statistical

sense. This predictive power is compellingly summarized by the conditional expectation function (CEF).

The CEF for a dependent variable Y_i, given a $K \times 1$ vector of covariates X_i (with elements x_{ki}), is the expectation, or population average, of Y_i, with X_i held fixed. The population average can be thought of as the mean in an infinitely large sample, or the average in a completely enumerated finite population. The CEF is written $E[Y_i|X_i]$ and is a function of X_i. Because X_i is random, the CEF is random, though sometimes we work with a particular value of the CEF, say $E[Y_i|X_i = 42]$, assuming 42 is a possible value for X_i. In chapter 2, we briefly considered the CEF $E[Y_i|D_i]$, where D_i is a zero-one variable. This CEF takes on two values, $E[Y_i|D_i = 1]$ and $E[Y_i|D_i = 0]$. Although this special case is important, we are most often interested in CEFs that are functions of many variables, conveniently subsumed in the vector X_i. For a specific value of X_i, say $X_i = x$, we write $E[Y_i|X_i = x]$. For continuous Y_i with conditional density $f_y(t|X_i = x)$ at $Y_i = t$, the CEF is

$$E[Y_i|X_i = x] = \int t f_y(t|X_i = x)dt.$$

If Y_i is discrete, $E[Y_i|X_i = x]$ equals the sum $\sum_t tP(Y_i = t|X_i = x)$, where $P(Y_i = t|X_i = x)$ is the conditional probability mass function for Y_i given $X_i = x$.

Expectation is a population concept. In practice, data usually come in the form of samples and rarely consist of an entire population. We therefore use samples to make inferences about the population. For example, the sample CEF is used to learn about the population CEF. This is necessary and important, but we postpone a discussion of the formal inference step taking us from sample to population until section 3.1.3. Our "population-first" approach to econometrics is motivated by the fact that we must define the objects of interest before we can use data to study them.[1]

[1]Examples of pedagogical writing using the "population-first" approach to econometrics include Chamberlain (1984), Goldberger (1991), and Manski (1991).

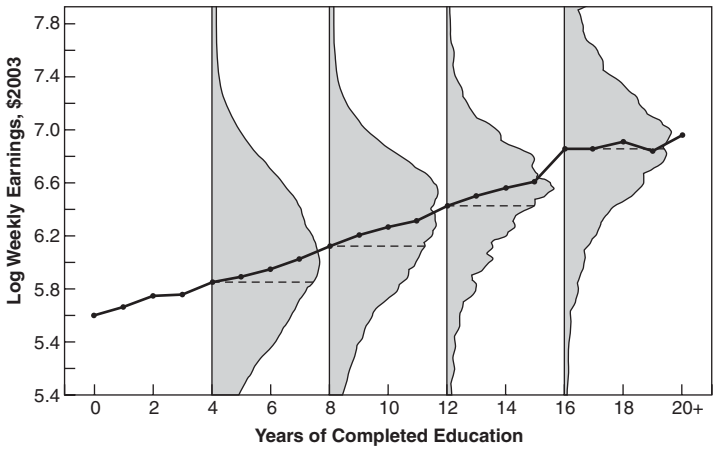

Figure 3.1.1 Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40–49 in the 1980 IPUMS 5 percent file.

Figure 3.1.1 plots the CEF of log weekly wages given schooling for a sample of middle-aged white men from the 1980 census. The distribution of earnings is also plotted for a few key values: 4, 8, 12, and 16 years of schooling. The CEF in the figure captures the fact that, notwithstanding the enormous variation individual circumstances, people with more schooling generally earn more. The average earnings gain associated with a year of schooling is typically about 10 percent.

An important complement to the CEF is the law of iterated expectations. This law says that an unconditional expectation can be written as the unconditional average of the CEF. In other words,

$$E[\mathrm{Y}_i] = E\{E[\mathrm{Y}_i|\mathrm{X}_i]\}, \tag{3.1.1}$$

where the outer expectation uses the distribution of X_i. Here is a proof of the law of iterated expectations for continuously distributed $(\mathrm{X}_i, \mathrm{Y}_i)$ with joint density $f_{xy}(u,t)$, where $f_y(t|\mathrm{X}_i = u)$ is the conditional distribution of Y_i given $\mathrm{X}_i = u$ and $g_y(t)$

and $g_x(u)$ are the marginal densities:

$$\begin{aligned}
E\{E[\mathrm{Y}_i|\mathrm{X}_i]\} &= \int E[\mathrm{Y}_i|\mathrm{X}_i = u]g_x(u)du \\
&= \int \left[\int t f_y(t|\mathrm{X}_i = u)dt\right] g_x(u)du \\
&= \int \int t f_y(t|\mathrm{X}_i = u)g_x(u)dudt \\
&= \int t\left[\int f_y(t|\mathrm{X}_i = u)g_x(u)du\right] dt \\
&= \int t\left[\int f_{xy}(u,t)du\right] dt \\
&= \int t g_y(t)dt = E[\mathrm{Y}_i].
\end{aligned}$$

The integrals in this derivation run over the possible values of X_i and Y_i (indexed by u and t). We've laid out these steps because the CEF and its properties are central to the rest of this chapter.[2]

The power of the law of iterated expectations comes from the way it breaks a random variable into two pieces, the CEF and a residual with special properties.

Theorem 3.1.1 *The CEF Decomposition Property.*

$$\mathrm{Y}_i = E[\mathrm{Y}_i|\mathrm{X}_i] + \varepsilon_i,$$

where (i) ε_i *is mean independent of* X_i*, that is,* $E[\varepsilon_i|\mathrm{X}_i] = 0$*, and therefore (ii)* ε_i *is uncorrelated with any function of* X_i*.*

Proof. (i) $E[\varepsilon_i|\mathrm{X}_i] = E[\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]|\mathrm{X}_i] = E[\mathrm{Y}_i|\mathrm{X}_i] - E[\mathrm{Y}_i|\mathrm{X}_i] = 0$. (ii) Let $h(\mathrm{X}_i)$ be any function of X_i. By the law of iterated expectations, $E[h(\mathrm{X}_i)\varepsilon_i] = E\{h(\mathrm{X}_i)E[\varepsilon_i|\mathrm{X}_i]\}$, and by mean independence, $E[\varepsilon_i|\mathrm{X}_i] = 0$.

[2] A simple example illustrates how the law of iterated expectations works: Average earnings in a population of men and women is the average for men times the proportion male in the population plus the average for women times the proportion female in the population.

This theorem says that any random variable Y_i can be decomposed into a piece that is "explained by X_i"—that is, the CEF—and a piece left over that is orthogonal to (i.e., uncorrelated with) any function of X_i.

The CEF is a good summary of the relationship between Y_i and X_i, for a number of reasons. First, we are used to thinking of averages as providing a representative value for a random variable. More formally, the CEF is the best predictor of Y_i given X_i in the sense that it solves a minimum mean squared error (MMSE) prediction problem. This CEF prediction property is a consequence of the CEF decomposition property:

Theorem 3.1.2 *The CEF Prediction Property.*
Let $m(X_i)$ be any function of X_i. The CEF solves

$$E[Y_i|X_i] = \underset{m(X_i)}{\arg\min}\, E[(Y_i - m(X_i))^2],$$

so it is the MMSE predictor of Y_i given X_i.

Proof. Write

$$\begin{aligned}(Y_i - m(X_i))^2 &= ((Y_i - E[Y_i|X_i]) + (E[Y_i|X_i] - m(X_i)))^2 \\ &= (Y_i - E[Y_i|X_i])^2 + 2(E[Y_i|X_i] - m(X_i)) \\ &\quad \times (Y_i - E[Y_i|X_i]) + (E[Y_i|X_i] - m(X_i))^2.\end{aligned}$$

The first term doesn't matter because it doesn't involve $m(X_i)$. The second term can be written $h(X_i)\varepsilon_i$, where $h(X_i) \equiv 2(E[Y_i|X_i] - m(X_i))$, and therefore has expectation zero by the CEF decomposition property. The last term is minimized at zero when $m(X_i)$ is the CEF.

A final property of the CEF, closely related to both the decomposition and prediction properties, is the analysis of variance (ANOVA) theorem:

Theorem 3.1.3 *The ANOVA Theorem.*

$$V(Y_i) = V(E[Y_i|X_i]) + E[V(Y_i|X_i)],$$

where $V(\cdot)$ denotes variance and $V(Y_i|X_i)$ is the conditional variance of Y_i given X_i.

Proof. The CEF decomposition property implies the variance of Y_i is the variance of the CEF plus the variance of the residual, $\varepsilon_i \equiv Y_i - E[Y_i|X_i]$, since ε_i and $E[Y_i|X_i]$ are uncorrelated. The variance of ε_i is

$$E[\varepsilon_i^2] = E[E[\varepsilon_i^2|X_i]] = E[V[Y_i|X_i]],$$

where $E[\varepsilon_i^2|X_i] = V[Y_i|X_i]$ because $\varepsilon_i \equiv Y_i - E[Y_i|X_i]$.

The two CEF properties and the ANOVA theorem may have a familiar ring. You might be used to seeing an ANOVA table in your regression output, for example. ANOVA is also important in research on inequality, where labor economists decompose changes in the income distribution into parts that can be accounted for by changes in worker characteristics and changes in what's left over after accounting for these factors (see, e.g., Autor, Katz, and Kearney, 2005). What may be unfamiliar is the fact that the CEF properties and ANOVA variance decomposition work in the population as well as in samples, and do not turn on the assumption of a linear CEF. In fact, the validity of linear regression as an empirical tool does not turn on linearity either.

3.1.2 *Linear Regression and the CEF*

So what's the regression you want to run? In our world, this question or one like it is heard almost every day. Regression estimates provide a valuable baseline for almost all empirical research because regression is tightly linked to the CEF, and the CEF provides a natural summary of empirical relationships. The link between regression functions—that is, the best-fitting line generated by minimizing expected squared errors—and the CEF can be explained in at least three ways. To lay out these explanations precisely, it helps to be precise about the regression function we have in mind. This section is concerned with the vector of population regression coefficients, defined as the solution to a population least squares problem. At this point we are not worried about causality. Rather,

we let the $\kappa \times 1$ regression coefficient vector β be defined by solving

$$\beta = \arg\min_b E[(\mathrm{Y}_i - \mathrm{X}_i'b)^2]. \tag{3.1.2}$$

Using the first-order condition,

$$E[\mathrm{X}_i(\mathrm{Y}_i - \mathrm{X}_i'b)] = 0,$$

the solution can be written $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$. Note that by construction, $E[\mathrm{X}_i(\mathrm{Y}_i - \mathrm{X}_i'\beta)] = 0$. In other words, the population residual, which we define as $\mathrm{Y}_i - \mathrm{X}_i'\beta = e_i$, is uncorrelated with the regressors, X_i. It bears emphasizing that this error term does not have a life of its own. It owes its existence and meaning to β. We return to this important point in the discussion of causal regression in section 3.2.

In the simple bivariate case where the regression vector includes only the single regressor, x_i, and a constant, the slope coefficient is $\beta_1 = \frac{Cov(\mathrm{Y}_i, x_i)}{V(x_i)}$, and the intercept is $\alpha = E[\mathrm{Y}_i] - \beta_1 E[\mathrm{X}_i]$. In the multivariate case, with more than one nonconstant regressor, the slope coefficient for the kth regressor is given below:

REGRESSION ANATOMY

$$\beta_k = \frac{Cov(\mathrm{Y}_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}, \tag{3.1.3}$$

where $\tilde{x}_{ki}$ is the residual from a regression of x_{ki} on all the other covariates.

In other words, $E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$ is the $\kappa \times 1$ vector with kth element $\frac{Cov(\mathrm{Y}_i, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}$. This important formula is said to describe the anatomy of a multivariate regression coefficient because it reveals much more than the matrix formula $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$. It shows us that each coefficient in a multivariate regression is the bivariate slope coefficient for the corresponding regressor after partialing out all the other covariates.

To verify the regression anatomy formula, substitute

$$Y_i = \alpha + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \cdots + \beta_K x_{Ki} + e_i$$

in the numerator of (3.1.3). Since $\tilde{x}_{ki}$ is a linear combination of the regressors, it is uncorrelated with e_i. Also, since $\tilde{x}_{ki}$ is a residual from a regression on all the other covariates in the model, it must be uncorrelated with these covariates. Finally, for the same reason, the covariance of $\tilde{x}_{ki}$ with x_{ki} is just the variance of $\tilde{x}_{ki}$. We therefore have $Cov(Y_i, \tilde{x}_{ki}) = \beta_k V(\tilde{x}_{ki})$.[3]

The regression anatomy formula is probably familiar to you from a regression or statistics course, perhaps with one twist: the regression coefficients defined in this section are not estimators; rather, they are nonstochastic features of the joint distribution of dependent and independent variables. This joint distribution is what you would observe if you had a complete enumeration of the population of interest (or knew the stochastic process generating the data). You probably don't have such information. Still, it's good empirical practice to think about what population parameters mean before worrying about how to estimate them.

Below we discuss three reasons why the vector of population regression coefficients might be of interest. These reasons can be summarized by saying that you should be interested in regression parameters if you are interested in the CEF.

[3]The regression anatomy formula is usually attributed to Frisch and Waugh (1933). You can also do regression anatomy this way:

$$\beta_k = \frac{Cov(\tilde{Y}_{ki}, \tilde{x}_{ki})}{V(\tilde{x}_{ki})},$$

where $\tilde{Y}_{ki}$ is the residual from a regression of Y_i on every covariate except x_{ki}. This works because the fitted values removed from $\tilde{Y}_{ki}$ are uncorrelated with $\tilde{x}_{ki}$. Often it's useful to plot $\tilde{Y}_{ki}$ against $\tilde{x}_{ki}$; the slope of the least squares fit in this scatterplot is the multivariate β_k, even though the plot is two-dimensional. Note, however, that it's not enough to partial the other covariates out of Y_i only. That is,

$$\frac{Cov(\tilde{Y}_{ki}, x_{ki})}{V(x_{ki})} = \left[\frac{Cov(\tilde{Y}_{ki}, \tilde{x}_{ki})}{V(\tilde{x}_{ki})}\right]\left[\frac{V(\tilde{x}_{ki})}{V(x_{ki})}\right] \neq \beta_k,$$

unless x_{ki} is uncorrelated with the other covariates.

Theorem 3.1.4 *The Linear CEF Theorem (Regression Justification I).*

Suppose the CEF is linear. Then the population regression function is it.

Proof. Suppose $E[\mathrm{Y}_i|\mathrm{X}_i] = \mathrm{X}_i'\beta^*$ for a $\kappa \times 1$ vector of coefficients, β^*. Recall that $E[\mathrm{X}_i(\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i])] = 0$ by the CEF decomposition property. Substitute using $E[\mathrm{Y}_i|\mathrm{X}_i] = \mathrm{X}_i'\beta^*$ to find that $\beta^* = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i] = \beta$.

The linear CEF theorem raises the question of what makes a CEF linear. The classic scenario is joint normality, that is, the vector $(\mathrm{Y}_i, \mathrm{X}_i')'$ has a multivariate normal distribution. This is the scenario considered by Galton (1886), father of regression, who was interested in the intergenerational link between normally distributed traits such as height and intelligence. The normal case is clearly of limited empirical relevance since regressors and dependent variables are often discrete, while normal distributions are continuous. Another linearity scenario arises when regression models are saturated. As reviewed in section 3.1.4, a saturated regression model has a separate parameter for every possible combination of values that the set of regressors can take on. For example a saturated regression model with two dummy covariates includes both covariates (with coefficients known as the main effects) and their product (known as an interaction term). Such models are inherently linear, a point we also discuss in section 3.1.4.

The following two reasons for focusing on regression are relevant when the linear CEF theorem does not apply.

Theorem 3.1.5 *The Best Linear Predictor Theorem (Regression Justification II).*

The function $\mathrm{X}_i'\beta$ is the best linear predictor of Y_i given X_i in a MMSE sense.

Proof. $\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$ solves the population least squares problem, (3.1.2).

In other words, just as the CEF, $E[\mathrm{Y}_i|\mathrm{X}_i]$, is the best (i.e., MMSE) predictor of Y_i given X_i in the class of *all* functions of

X_i, the population regression function is the best we can do in the class of *linear* functions.

Theorem 3.1.6 *The Regression CEF Theorem (Regression Justification III).*

The function $X_i'\beta$ provides the MMSE linear approximation to $E[Y_i|X_i]$, that is,

$$\beta = \arg\min_b E\{(E[Y_i|X_i] - X_i'b)^2\}. \tag{3.1.4}$$

Proof. Start by observing that β solves (3.1.2). Write

$$\begin{aligned}(Y_i - X_i'b)^2 &= \{(Y_i - E[Y_i|X_i]) + (E[Y_i|X_i] - X_i'b)\}^2 \\ &= (Y_i - E[Y_i|X_i])^2 + (E[Y_i|X_i] - X_i'b)^2 \\ &\quad + 2(Y_i - E[Y_i|X_i])(E[Y_i|X_i] - X_i'b).\end{aligned}$$

The first term doesn't involve b and the last term has expectation zero by the CEF decomposition property (ii). The CEF approximation problem, (3.1.4), is therefore the same as the population least squares problem, (3.1.2).

These two theorems give us two more ways to view regression. Regression provides the best linear predictor for the dependent variable in the same way that the CEF is the best unrestricted predictor of the dependent variable. On the other hand, if we prefer to think about approximating $E[Y_i|X_i]$, as opposed to predicting Y_i, the regression CEF theorem tells us that even if the CEF is nonlinear, regression provides the best linear approximation to it.

The regression CEF theorem is our favorite way to motivate regression. The statement that regression approximates the CEF lines up with our view of empirical work as an effort to describe the essential features of statistical relationships without necessarily trying to pin them down exactly. The linear CEF theorem is for special cases only. The best linear predictor theorem is satisfyingly general, but seems to encourage an overly clinical view of empirical research. We're not really interested in predicting *individual* Y_i; it's the *distribution* of Y_i that we care about.

Figure 3.1.2 Regression threads the CEF of average weekly wages given schooling (dots = CEF; dashes = regression line).

Figure 3.1.2 illustrates the CEF approximation property for the same schooling CEF plotted in figure 3.1.1. The regression line fits the somewhat bumpy and nonlinear CEF as if we were estimating a model for $E[\mathrm{Y}_i|\mathrm{X}_i]$ instead of a model for Y_i. In fact, that is exactly what's going on. An implication of the regression CEF theorem is that regression coefficients can be obtained by using $E[\mathrm{Y}_i|\mathrm{X}_i]$ as a dependent variable instead of Y_i itself. To see this, suppose that X_i is a discrete random variable with probability mass function $g_x(u)$. Then

$$E\{(E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'b)^2\} = \sum_u (E[\mathrm{Y}_i|\mathrm{X}_i = u] - u'b)^2 g_x(u).$$

This means that β can be constructed from the weighted least squares (WLS) regression of $E[\mathrm{Y}_i|\mathrm{X}_i = u]$ on u, where u runs over the values taken on by X_i. The weights are given by the distribution of X_i, that is, $g_x(u)$. An even simpler way to see this is to iterate expectations in the formula for β:

$$\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i] = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i E(\mathrm{Y}_i|\mathrm{X}_i)]. \quad (3.1.5)$$

The CEF or grouped data version of the regression formula is of practical use when working on a project that precludes the analysis of microdata. For example, Angrist (1998) used grouped data to study the effect of voluntary military service on earnings later in life. One of the estimation strategies used in this project regresses civilian earnings on a dummy for veteran status, along with personal characteristics and the variables used by the military to screen soldiers. The earnings data come from the U.S. Social Security system, but Social Security earnings records cannot be released to the public. Instead of individual earnings, Angrist worked with average earnings conditional on race, sex, test scores, education, and veteran status.

To illustrate the grouped data approach to regression, we estimated the schooling coefficient in a wage equation using 21 conditional means, the sample CEF of earnings given schooling. As the Stata output reproduced in Figure 3.1.3 shows, a grouped data regression, weighted by the number of individuals at each schooling level in the sample, produces coefficients identical to those generated using the underlying microdata sample with hundreds of thousands of observations. Note, however, that the standard errors from the grouped regression do not measure the asymptotic sampling variance of the slope estimate in repeated micro-data samples; for that you need an estimate of the variance of $\mathrm{Y}_i - \mathrm{X}_i'\beta$. This variance depends on the microdata, in particular the second moments of $W_i \equiv [\mathrm{Y}_i \ \mathrm{X}_i']'$, a point we elaborate on in the next section.

3.1.3 *Asymptotic OLS Inference*

In practice, we don't usually know what the CEF or the population regression vector is. We therefore draw statistical inferences about these quantities using samples. Statistical inference is what much of traditional econometrics is about. Although this material is covered in any econometrics text, we don't want to skip the inference step completely. A review of basic asymptotic theory allows us to highlight the important fact that the process of statistical inference is distinct from the

A - Individual-level data

```
. regress earnings school, robust

      Source |       SS       df       MS        Number of obs =  409435
-------------+------------------------------     F(  1,409433) =49118.25
       Model | 22631.4793      1  22631.4793     Prob > F      =  0.0000
    Residual |  188648.31 409433  .460755019     R-squared     =  0.1071
-------------+------------------------------     Adj R-squared =  0.1071
       Total | 211279.789 409434   .51602893     Root MSE      =  .67879
-------------+----------------------------------------------------------------
             |               Robust                       Old Fashioned
    earnings |      Coef.   Std. Err.      t           Std. Err.       t
-------------+----------------------------------------------------------------
      school |   .0674387   .0003447   195.63          .0003043   221.63
      const. |   5.835761   .0045507  1282.39          .0040043  1457.38
------------------------------------------------------------------------------
```

B - Means by years of schooling

```
. regress average_earnings school [aweight=count], robust
(sum of wgt is   4.0944e+05)

      Source |       SS       df       MS        Number of obs =      21
-------------+------------------------------     F(  1,    19) =  540.31
       Model |  1.16077332     1  1.16077332     Prob > F      =  0.0000
    Residual |  .040818796    19  .002148358     R-squared     =  0.9660
-------------+------------------------------     Adj R-squared =  0.9642
       Total |  1.20159212    20  .060079606     Root MSE      =  .04635
-------------+----------------------------------------------------------------
     average |               Robust                       Old Fashioned
   _earnings |      Coef.   Std. Err.      t           Std. Err.       t
-------------+----------------------------------------------------------------
      school |   .0674387   .0040352    16.71          .0029013     23.24
      const. |   5.835761   .0399452   146.09          .0381792    152.85
------------------------------------------------------------------------------
```

Figure 3.1.3 Microdata and grouped data estimates of the returns to schooling, from Stata regression output. *Source*: 1980 Census—IPUMS, 5 percent sample. The sample includes white men, age 40–49. Robust standard errors are heteroskedasticity consistent. Panel A uses individual-level microdata. Panel B uses earnings averaged by years of schooling.

question of how a particular set of regression estimates should be interpreted. Whatever a regression coefficient may mean, it has a sampling distribution that is easy to describe and use for statistical inference.[4]

[4]The discussion of asymptotic OLS inference in this section is largely a condensation of material in Chamberlain (1984). Important pitfalls and problems with asymptotic theory are covered in the last chapter.

We are interested in the distribution of the sample analog of

$$\beta = E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{Y}_i]$$

in repeated samples. Suppose the vector $W_i \equiv [\mathrm{Y}_i\ \mathrm{X}_i']'$ is independently and identically distributed in a sample of size N. A natural estimator of the first population moment, $E[W_i]$, is the sum, $\frac{1}{N}\sum_{i=1}^{N} W_i$. By the law of large numbers, this vector of sample moments gets arbitrarily close to the corresponding vector of population moments as the sample size grows. We might similarly consider higher-order moments of the elements of W_i, for example the matrix of second moments, $E[W_iW_i']$, with sample analog $\frac{1}{N}\sum_{i=1}^{N} W_iW_i'$. Following this principle, the method of moments estimator of β replaces each expectation by a sum. This logic leads to the ordinary least squares (OLS) estimator

$$\hat{\beta} = \left[\sum_i \mathrm{X}_i\mathrm{X}_i'\right]^{-1} \sum_i \mathrm{X}_i\mathrm{Y}_i.$$

Although we derived $\hat{\beta}$ as a method of moments estimator, it is called the OLS estimator of β because it solves the sample analog of the least squares problem described at the beginning of section 3.1.2.[5]

The asymptotic sampling distribution of $\hat{\beta}$ depends solely on the definition of the estimand (i.e., the nature of the thing we're trying to estimate, β) and the assumption that the data constitute a random sample. Before deriving this distribution, it helps to summarize the general asymptotic distribution theory that covers our needs. This basic theory can be stated mostly in words. For the purposes of these statements, we assume the reader is familiar with the core terms and concepts of statistical theory—moments, mathematical expectation, probability

[5]Econometricians like to use matrices because the notation is so compact. Sometimes (not very often) we do too. Suppose X is the matrix whose rows are given by X_i' and y is the vector with elements Y_i, for $i = 1, \ldots, N$. The sample moment matrix $\frac{1}{N}\sum X_iX_i'$ is $X'X/N$ and the sample moment vector $\frac{1}{N}\sum X_iy_i$ is $X'y/N$. Then we can write $\hat{\beta} = (X'X)^{-1}X'y$, a widely used matrix formula.

limits, and asymptotic distributions. For definitions of these terms and a formal mathematical statement of the theoretical propositions given below, see Knight (2000).

THE LAW OF LARGE NUMBERS Sample moments converge in probability to the corresponding population moments. In other words, the probability that the sample mean is close to the population mean can be made as high as you like by taking a large enough sample.

THE CENTRAL LIMIT THEOREM Sample moments are asymptotically normally distributed (after subtracting the corresponding population moment and multiplying by the square root of the sample size). The asymptotic covariance matrix is given by the variance of the underlying random variable. In other words, in large enough samples, appropriately normalized sample moments are approximately normally distributed.

SLUTSKY'S THEOREM

1. Consider the sum of two random variables, one of which converges in distribution (in other words, has an asymptotic distribution) and the other converges in probability to a constant: the asymptotic distribution of this sum is unaffected by replacing the one that converges to a constant by this constant. Formally, let a_N be a statistic with an asymptotic distribution and let b_N be a statistic with probability limit b. Then $a_N + b_N$ and $a_N + b$ have the same asymptotic distribution.
2. Consider the product of two random variables, one of which converges in distribution and the other converges in probability to a constant: the asymptotic distribution of this product is unaffected by replacing the one that converges to a constant by this constant. Formally, let a_N be a statistic with an asymptotic distribution and let b_N be a statistic with probability limit b. Then $a_N b_N$ and $a_N b$ have the same asymptotic distribution.

THE CONTINUOUS MAPPING THEOREM Probability limits pass through continuous functions. For example, the probability

limit of any continuous function of a sample moment is the function evaluated at the corresponding population moment. Formally, the probability limit of $h(b_N)$ is $h(b)$, where *plim* $b_N = b$ and $h(\cdot)$ is continuous at b.

THE DELTA METHOD Consider a vector-valued random variable that is asymptotically normally distributed. Continuously differentiable scalar functions of this random variable are also asymptotically normally distributed, with covariance matrix given by a quadratic form with the covariance matrix of the random variable on the inside and the gradient of the function evaluated at the probability limit of the random variable on the outside.[6] Formally, the asymptotic distribution of $h(b_N)$ is normal with covariance matrix $\nabla h(b)'\Omega\nabla h(b)$, where *plim* $b_N = b$, $h(\cdot)$ is continuously differentiable at b with gradient $\nabla h(b)$, and b_N has asymptotic covariance matrix Ω.[7]

We can use these results to derive the asymptotic distribution of $\hat{\beta}$ in two ways. A conceptually straightforward but somewhat inelegant approach is to use the delta method: $\hat{\beta}$ is a function of sample moments, and is therefore asymptotically normally distributed. It remains only to find the covariance matrix of the asymptotic distribution from the gradient of this function. (Note that consistency of $\hat{\beta}$ comes immediately from the continuous mapping theorem).[8] An easier and more instructive derivation uses the Slutsky and central limit theorems. Note first that we can write

$$\mathrm{Y}_i = \mathrm{X}_i'\beta + [\mathrm{Y}_i - \mathrm{X}_i'\beta] \equiv \mathrm{X}_i'\beta + e_i, \tag{3.1.6}$$

where the residual e_i is defined as the difference between the dependent variable and the population regression function, as

[6] A quadratic form is a matrix-weighted sum of squares. Suppose v is an $N \times 1$ vector and M is an $N \times N$ matrix. A quadratic form in v is $v'Mv$. If M is an $N \times N$ diagonal matrix with diagonal elements m_i, then $v'Mv = \sum_i m_i v_i^2$.

[7] For a derivation of the delta method formula using the Slutsky and continuous mapping theorems, see Knight (2000, pp. 120–121). We say "the asymptotic distribution of $h(b_N)$," but we really mean the asymptotic distribution of $\sqrt{N}(h(b_N) - h(b))$.

[8] An estimator is said to be *consistent* when it converges in probability to the target parameter.

before. In other words, $E[X_i e_i] = 0$ is a consequence of $\beta = E[X_i\, X_i']^{-1}E[X_i Y_i]$ and $e_i = Y_i - X_i'\beta$, and not an assumption about an underlying economic relation.[9]

Substituting the identity (3.1.6) for Y_i in the formula for $\hat{\beta}$, we have

$$\hat{\beta} = \beta + \left[\sum X_i X_i'\right]^{-1} \sum X_i e_i.$$

The asymptotic distribution of $\hat{\beta}$ is the asymptotic distribution of $\sqrt{N}(\hat{\beta} - \beta) = N[\sum X_i X_i']^{-1} \frac{1}{\sqrt{N}} \sum X_i e_i$. By the Slutsky theorem, this has the same asymptotic distribution as $E[X_i X_i']^{-1} \frac{1}{\sqrt{N}} \sum X_i e_i$. Since $E[X_i e_i] = 0$, $\frac{1}{\sqrt{N}} \sum X_i e_i$ is a root-N normalized and centered sample moment. By the central limit theorem, this is asymptotically normally distributed with mean zero and covariance matrix $E[X_i X_i' e_i^2]$, since this matrix of fourth moments is the covariance matrix of $X_i e_i$. Therefore, $\hat{\beta}$ has an asymptotic normal distribution with probability limit β and covariance matrix

$$E[X_i X_i']^{-1} E[X_i X_i' e_i^2] E[X_i X_i']^{-1}. \tag{3.1.7}$$

The theoretical standard errors used to construct t-statistics are the square roots of the diagonal elements of (3.1.7). In practice these standard errors are estimated by substituting sums for expectations and using the estimated residuals, $\hat{e}_i = Y_i - X_i'\hat{\beta}$ to form the empirical fourth moment matrix, $\sum[X_i X_i \hat{e}_i^2]/N$.

Asymptotic standard errors computed in this way are known as heteroskedasticity-consistent standard errors, White (1980a) standard errors, or Eicker-White standard errors, in recognition of Eicker's (1967) derivation. They are also known as "robust" standard errors (e.g., in Stata). These standard errors are said to be robust because, in large enough samples, they provide accurate hypothesis tests and confidence intervals given minimal assumptions about the data and model. In particular, our derivation of the limiting distribution makes

[9] Residuals defined in this way are not necessarily mean independent of X_i; for mean independence, we need a linear CEF.

no assumptions other than those needed to ensure that basic statistical results like the central limit theorem go through. Robust standard errors are not, however, the standard errors that you get by default from packaged software. Default standard errors are derived under a homoskedasticity assumption, specifically, that $E[e_i^2|\mathrm{X}_i] = \sigma^2$, a constant. Given this assumption, we have

$$E[\mathrm{X}_i\mathrm{X}_i'e_i^2] = E(\mathrm{X}_i\mathrm{X}_i'E[e_i^2|\mathrm{X}_i]) = \sigma^2 E[\mathrm{X}_i\mathrm{X}_i'],$$

by iterating expectations. The asymptotic covariance matrix of $\hat{\beta}$ then simplifies to

$$\begin{aligned} &E[\mathrm{X}_i\mathrm{X}_i']^{-1}E[\mathrm{X}_i\mathrm{X}_i'e_i^2]E[\mathrm{X}_i\mathrm{X}_i']^{-1} \\ &\quad = E[\mathrm{X}_i\mathrm{X}_i']^{-1}\sigma^2 E[\mathrm{X}_i\mathrm{X}_i']E[\mathrm{X}_i\mathrm{X}_i]^{-1} \\ &\quad = \sigma^2 E[\mathrm{X}_i\mathrm{X}_i']^{-1}. \end{aligned} \tag{3.1.8}$$

The diagonal elements of (3.1.8) are what SAS or Stata report unless you request otherwise.

Our view of regression as an approximation to the CEF makes heteroskedasticity seem natural. If the CEF is nonlinear and you use a linear model to approximate it, then the quality of fit between the regression line and the CEF will vary with X_i. Hence, the residuals will be larger, on average, at values of X_i where the fit is poorer. Even if you are prepared to assume that the conditional variance of Y_i given X_i is constant, the fact that the CEF is nonlinear means that $E[(\mathrm{Y}_i - \mathrm{X}_i'\beta)^2|\mathrm{X}_i]$ will vary with X_i. To see this, note that

$$\begin{aligned} &E[(\mathrm{Y}_i - \mathrm{X}_i'\beta)^2|\mathrm{X}_i] \\ &\quad = E\{[(\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) + (E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'\beta)]^2|\mathrm{X}_i\} \\ &\quad = V[\mathrm{Y}_i|\mathrm{X}_i] + (E[\mathrm{Y}_i|\mathrm{X}_i] - \mathrm{X}_i'\beta)^2. \end{aligned} \tag{3.1.9}$$

Therefore, even if $V[\mathrm{Y}_i|\mathrm{X}_i]$ is constant, the residual variance increases with the square of the gap between the regression line and the CEF, a fact noted in White (1980b).[10]

[10]The cross-product term resulting from an expansion of the squared term in the middle of (3.1.9) is zero because $\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]$ is mean independent of X_i.

In the same spirit, it's also worth noting that while a linear CEF makes homoskedasticity possible, this is not a sufficient condition for homoskedasticity. Our favorite example in this context is the linear probability model (LPM). A linear probability model is any regression where the dependent variable is zero-one, that is, a dummy variable such as an indicator for labor force participation. Suppose the regression model is saturated, so the CEF given regressors is linear. Because the CEF is linear, the residual variance is also the conditional variance, $V[\mathrm{Y}_i|\mathrm{X}_i]$. But the dependent variable is a Bernoulli trial with conditional variance $P[\mathrm{Y}_i = 1|\mathrm{X}_i](1 - P[\mathrm{Y}_i = 1|\mathrm{X}_i])$. We conclude that LPM residuals are necessarily heteroskedastic unless the only regressor is a constant.

These points of principle notwithstanding, as an empirical matter, heteroskedasticity may matter little. In the microdata schooling regression depicted in figure 3.1.3, the robust standard error is .0003447, while the old-fashioned standard error is .0003043, not much smaller. The standard errors from the grouped data regression, which are necessarily heteroskedastic if group sizes differ, change somewhat more; compare the .004 robust standard to the .0029 conventional standard error. Based on our experience, these differences are typical. If heteroskedasticity matters a lot, say, more than a 30 percent increase or any marked decrease in standard errors, you should worry about possible programming errors or other problems. For example, robust standard errors below conventional may be a sign of finite-sample bias in the robust calculation.

Finally, a brief note on the textbook approach to inference that you might have seen elsewhere. Traditional econometric inference begins with stronger assumptions than those we have invoked in this section. The traditional set-up, sometimes called a classical normal regression model, postulates: fixed (non-stochastic) regressors, a linear CEF, normally distributed errors, and homoskedasticity (see, e.g., Goldberger, 1991). These stronger assumptions give us two things: (1) unbiasedness of the OLS estimator, (2) a formula for the sampling variance of the OLS estimator that is valid in small as well as large samples. Unbiasedness of the OLS estimators means that $E[\hat{\beta}] = \beta$, a property that holds in a sample of any size and is

stronger than consistency, which means only that we can expect $\hat{\beta}$ to be close to β in large samples. It's easy to see when and why we get unbiasedness. In general,

$$E[\hat{\beta}] = \beta + E\left\{\left[\sum \mathrm{X}_i \mathrm{X}_i'\right]^{-1} \sum \mathrm{X}_i e_i\right\}.$$

If the regressors are nonrandom (fixed in repeated samples) the expectation passes through and we have unbiasedness because $E[e_i] = 0$. Otherwise, with random regressors, we can iterate expectations and get unbiasedness if $E[e_i|\mathrm{X}_i] = 0$. This is true when the CEF is linear, but not in our more general "agnostic regression" framework.

The variance formula obtained under classical assumptions is the same as the large-sample formula under homoskedasticity but—provided the strong classical assumptions are valid—this formula holds in a sample of any size. We've chosen to start with the asymptotic approach to inference because modern empirical work typically leans heavily on the large-sample theory that lies behind robust variance formulas. The payoff is valid inference under weak assumptions, in particular, a framework that makes sense for our less-than-literal approach to regression models. On the other hand, the large-sample approach is not without its dangers, a point we return to in the discussion of inference in chapter 8 and in the discussion of instrumental variables in chapter 4.

3.1.4 *Saturated Models, Main Effects, and Other Regression Talk*

We often discuss regression models using terms like *saturated* and *main effects*. These terms originate in an experimentalist tradition that uses regression to model the effects of discrete treatment-type variables. This language is now used more widely in many fields, however, including applied econometrics. For readers unfamiliar with these terms, this section provides a brief review.

Saturated regression models are regression models with discrete explanatory variables, where the model includes a separate parameter for all possible values taken on by the

explanatory variables. For example, when working with a single explanatory variable indicating whether a worker is a college graduate, the model is saturated by including a single dummy for college graduates and a constant. We can also saturate when the regressor takes on many values. Suppose, for example, that $S_i = 0, 1, 2, \ldots, \tau$. A saturated regression model for S_i is

$$Y_i = \alpha + \beta_1 d_{1i} + \beta_2 d_{2i} + \cdots + \beta_\tau d_{\tau i} + \varepsilon_i,$$

where $d_{ji} = 1[S_i = j]$ is a dummy variable indicating schooling level j, and β_j is said to be the jth-level schooling *effect*.[11] Note that

$$\beta_j = E[Y_i|S_i = j] - E[Y_i|S_i = 0],$$

while $\alpha = E[Y_i|S_i = 0]$. In practice, you can pick any value of S_i for the reference group; a regression model is saturated as long as it has one parameter for every possible j in $E[Y_i|S_i = j]$. Saturated regression models fit the CEF perfectly because the CEF is a linear function of the dummy regressors used to saturate. This is an important special case of the linear CEF theorem.

If there are two explanatory variables—say, one dummy indicating college graduates and one dummy indicating sex—the model is saturated by including these two dummies, their product, and a constant. The coefficients on the dummies are known as main effects, while the product is called an *interaction term*. This is not the only saturated parameterization; any set of indicators (dummies) that can be used to identify each value taken on by all covariates produces a saturated model. For example, an alternative saturated model includes dummies for male college graduates, male nongraduates, female college graduates, and female nongraduates, but no intercept.

Here's some notation to make this more concrete. Let x_{1i} indicate college graduates and x_{2i} indicate women. The CEF

[11] We use the notation $1[S_i = j]$ to denote the indicator function, in this case a function that creates a dummy variable switched on when $S_i = j$.

given x_{1i} and x_{2i} takes on four values:

$$E[Y_i|x_{1i} = 0, x_{2i} = 0],$$
$$E[Y_i|x_{1i} = 1, x_{2i} = 0],$$
$$E[Y_i|x_{1i} = 0, x_{2i} = 1],$$
$$E[Y_i|x_{1i} = 1, x_{2i} = 1].$$

We can label these using the following scheme:

$$E[Y_i|x_{1i} = 0, x_{2i} = 0] = \alpha$$
$$E[Y_i|x_{1i} = 1, x_{2i} = 0] = \alpha + \beta_1$$
$$E[Y_i|x_{1i} = 0, x_{2i} = 1] = \alpha + \gamma$$
$$E[Y_i|x_{1i} = 1, x_{2i} = 1] = \alpha + \beta_1 + \gamma + \delta_1.$$

Since there are four Greek letters and the CEF takes on four values, this parameterization does not restrict the CEF. It can be written in terms of Greek letters as

$$E[Y_i|x_{1i}, x_{2i}] = \alpha + \beta_1 x_{1i} + \gamma x_{2i} + \delta_1(x_{1i}x_{2i}),$$

a parameterization with two main effects and one interaction term.[12] The saturated regression equation becomes

$$Y_i = \alpha + \beta_1 x_{1i} + \gamma x_{2i} + \delta_1(x_{1i}x_{2i}) + \varepsilon_i.$$

We can combine the multivalued schooling variable with sex to produce a saturated model that has τ main effects for schooling, one main effect for sex, and τ sex-schooling interactions:

$$Y_i = \alpha + \sum_{j=1}^{\tau} \beta_j d_{ji} + \gamma x_{2i} + \sum_{j=1}^{\tau} \delta_j (d_{ji}x_{2i}) + \varepsilon_i. \qquad (3.1.10)$$

The coefficients on the interaction terms, δ_j, tell us how each of the schooling effects differ by sex. The CEF in this case

[12] With a third dummy variable in the model, say x_{3i}, a saturated model includes three main effects, three second-order interaction terms $\{x_{1i}x_{2i}, x_{1i}x_{3i}, x_{2i}x_{3i}\}$, and one third-order term, $x_{1i}x_{2i}x_{3i}$.