3.1 Regression Fundamentals

The end of the previous chapter introduces 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 cases, 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. This chapter begins by reviewing these properties, which include:

(i) the intimate connection between the population regression function and the conditional expectation function

(ii) how and why regression coefficients change as covariates are added or removed from the model

(iii) the close link between regression and other "control strategies" such as matching

(iv) 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. Such 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 average 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 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 will 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 $\text{K}\times 1$ 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 $E\left[\text{Y}_i|\text{X}_i\right]$ 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 $E[\text{Y}_i|\text{X}_i{=}42]$, assuming 42 is a possible value for X_i. In Chapter 2, we briefly considered the CEF $E[\text{Y}_i|\text{D}_i]$, where D_i is a zero-one variable. This CEF takes on two values, $E[\text{Y}_i|\text{D}_i = 1]$ and $E[\text{Y}_i|\text{D}_i = 0]$. 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 $\text{X}_i = x$, we write $E\left[\text{Y}_i|\text{X}_i = x\right]$. For continuous Y_i with conditional density $f_y\left(\cdot|\text{X}_i = x\right)$, the CEF is

$$E\left[\text{Y}_i|\text{X}_i = x\right] = \int t f_y\left(t|\text{X}_i = x\right) dt.$$

If Y_i is discrete, $E\left[\text{Y}_i|\text{X}_i = x\right]$ equals the sum $\sum_t t f_y\left(t|\text{X}_i = x\right)$.

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 always necessary 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]

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—the enormous variation individual circumstances

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

notwithstanding—people with more schooling generally earn more, on average. The average earnings gain associated with a year of schooling is typically about 10 percent.

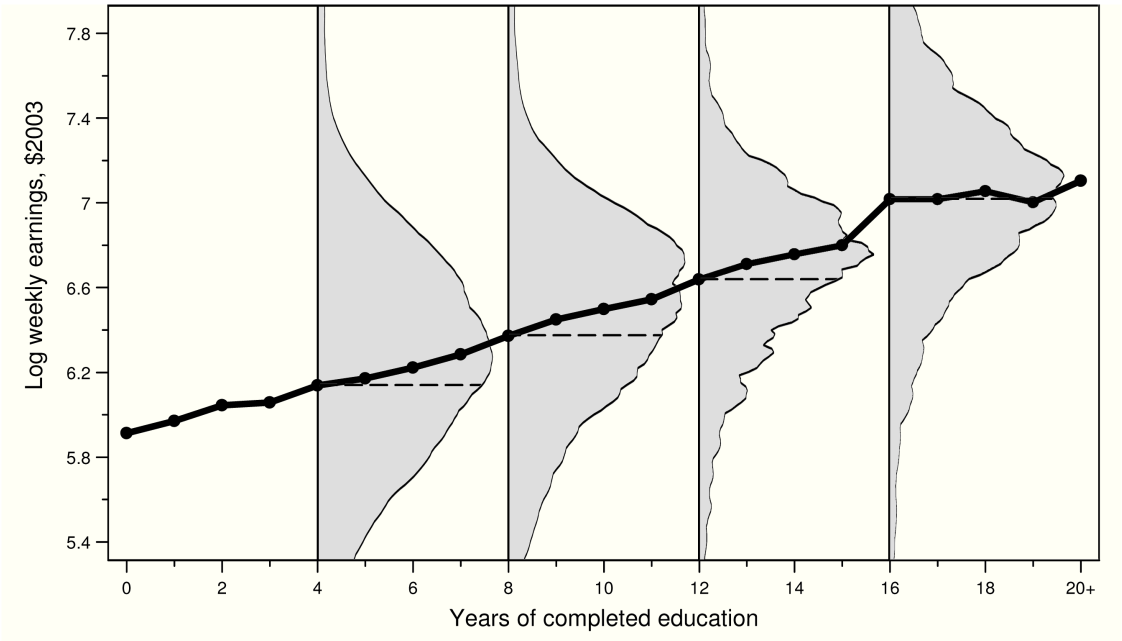

Figure 3.1.1: Raw data and the CEF of average log weekly wages given schooling. The sample includes white men aged 40-49 in the 1980 IPUMS 5 percent file.

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

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

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

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

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.

The power of the law of iterated expectations comes from the way it breaks a random variable into two pieces.

Theorem 3.1.1 *The CEF-Decomposition Property*

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

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

Proof. (i) $E[\varepsilon_i|\mathrm{X}_i] = E[\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i] \mid \mathrm{X}_i] = E[\mathrm{Y}_i|\mathrm{X}_i] - E[\mathrm{Y}_i|\mathrm{X}_i] = 0$;(ii) This follows from (i): 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, $E[\varepsilon_i|\mathrm{X}_i] = 0$. ■

This theorem says that any random variable, Y_i, can be decomposed into a piece that's "explained by X_i", i.e., the CEF, and a piece left over which 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 $m(\mathrm{X}_i)$ be any function of X_i. The CEF solves

$$E[\mathrm{Y}_i|\mathrm{X}_i] = \underset{m(\mathrm{X}_i)}{\arg\min} E\left[(\mathrm{Y}_i - m(\mathrm{X}_i))^2\right],$$

so it is the MMSE predictor of Y_i given X_i.

Proof. Write

$$\begin{aligned}(\mathrm{Y}_i - m(\mathrm{X}_i))^2 &= ((\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) + (E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i)))^2 \\ &= (\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i])^2 + 2(E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i))(\mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]) \\ &\quad + (E[\mathrm{Y}_i|\mathrm{X}_i] - m(\mathrm{X}_i))^2\end{aligned}$$

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

A final property of the CEF, closely related to both the CEF decomposition and prediction properties, is the Analysis-of-Variance (ANOVA) Theorem:

Theorem 3.1.3 *The ANOVA Theorem*

$$V(\mathrm{Y}_i) = V(E[\mathrm{Y}_i|\mathrm{X}_i]) + E[V(\mathrm{Y}_i|\mathrm{X}_i)]$$

where $V(\cdot)$ denotes variance and $V(\mathrm{Y}_i|\mathrm{X}_i)$ 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, $\varepsilon_i \equiv \mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]$ since ε_i and $E[\mathrm{Y}_i|\mathrm{X}_i]$ are uncorrelated. The variance of ε_i is

$$E\left[\varepsilon_i^2\right] = E\left[E\left[\varepsilon_i^2|\mathrm{X}_i\right]\right] = E[V[\mathrm{Y}_i|\mathrm{X}_i]]$$

where $E\left[\varepsilon_i^2|\mathrm{X}_i\right] = V[\mathrm{Y}_i|\mathrm{X}_i]$ because $\varepsilon_i \equiv \mathrm{Y}_i - E[\mathrm{Y}_i|\mathrm{X}_i]$. ■

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 – i.e., the best-fitting line generated by minimizing expected squared errors – and the CEF can be explained in at least 3 ways. To lay out these explanations precisely, it helps to be precise about the regression function we