6 REGRESSION ANALYSIS

An Introduction to Regression Analysis

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the "statistical significance" of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship.

Regression techniques have long been central to the field of economic statistics ("econometrics"). Increasingly, they have become important to lawyers and legal policymakers as well. Regression has been offered as evidence of liability under Title VII of the Civil Rights Act of 1964 as evidence of racial bias in death penalty litigation, as evidence of damages in contract actions, as evidence of violations under the Voting Rights Act, and as evidence of damage in antitrust litigation, among other things.

In this course notes, we will have an overview of the most basic techniques of regression analysis—how they work, what they assume, and how they may go awry when key assumptions do not hold. To make the discussion concrete, I will employ a series of illustrations involving a hypothetical analysis of the factors that determine individual earnings in the labor market. The illustrations will have a legal flavor in the latter part of the lecture, where they will incorporate the possibility that earnings are impermissibly influenced by gender in violation of the federal civil rights laws. Also, of necessity, there are many important topics that I omit, including simultaneous equation models and generalized least squares. The lecture is limited to the assumptions, mechanics, and common difficulties with a single equation, ordinary least squares regression.

What is Regression?

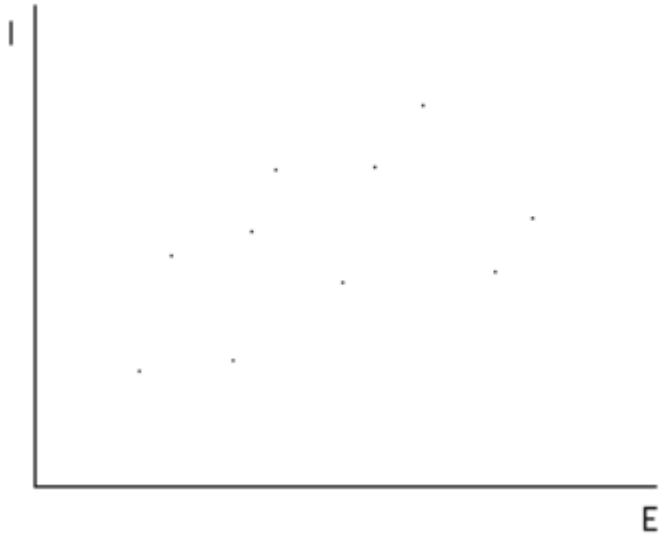
For purposes of illustration, suppose that we wish to identify and quantify the factors that determine earnings in the labor market. A moment's reflection suggests a myriad of factors that are associated with variations in earnings across individuals—occupation, age, experience, educational attainment, motivation, and innate ability come to mind, perhaps along with factors such as race and gender that can be of particular concern to lawyers. For the time being, let us restrict attention to a single factor—call it education. Regression analysis with a single explanatory variable is termed "simple regression."

Simple Regression

In reality, any effort to quantify the effects of education upon earnings without careful attention to the other factors that affect earnings could create serious statistical difficulties (termed "omitted variables bias"), which I will discuss later. But for now, let us assume away this problem. We also assume, again quite unrealistically, that "education" can be measured by a single attribute—years of schooling. We thus suppress the fact that a given number of years in school may represent widely varying academic programs.

At the outset of any regression study, one formulates some hypothesis about the relationship between the variables of interest, here, education and earnings. Common experience suggests that better-educated people tend to make more money. It further suggests that the causal relation likely runs from education to earnings rather than the other way around. Thus, the tentative hypothesis is that higher levels of education cause higher levels of earnings, other things being equal.

To investigate this hypothesis, imagine that we gather data on education and earnings for various individuals. Let E denote education in years of schooling for each individual, and let I denote that individual's earnings in dollars per year. We can plot this information for all of the individuals in the sample using a two-dimensional diagram, conventionally termed a "scatter" diagram. Each point in the diagram represents an individual in the sample.



The diagram indeed suggests that higher values of E tend to yield higher values of I, but the relationship is not perfect it seems that knowledge of E does not suffice for an entirely accurate prediction about I. We can then deduce either that the effect of education on earnings differs across individuals, or that factors other than education influence earnings. Regression analysis ordinarily embraces the latter explanation. Thus, pending discussion below of omitted variables bias, we now hypothesize that earnings for each individual are determined by education and by an aggregation of omitted factors that we term "noise."

To refine the hypothesis further, it is natural to suppose that people in the labor force with no education nevertheless make some positive amount of money and that education increases earnings above this baseline. We might also suppose that education affects income in a "linear" fashion—that is, each additional year of schooling adds the same amount to income. This linearity assumption is common in regression studies but is by no means essential to the application of the technique, and can be relaxed where the investigator has reason to suppose a priori that the relationship in question is nonlinear.

Then, the hypothesized relationship between education and earnings may be written

$I = \alpha + \beta E + \varepsilon$

where
α = a constant amount (what one earns with zero education);
β = the effect in dollars of an additional year of schooling on income hypothesized to be positive; and
ε = the "noise" term reflecting other factors that influence earnings.

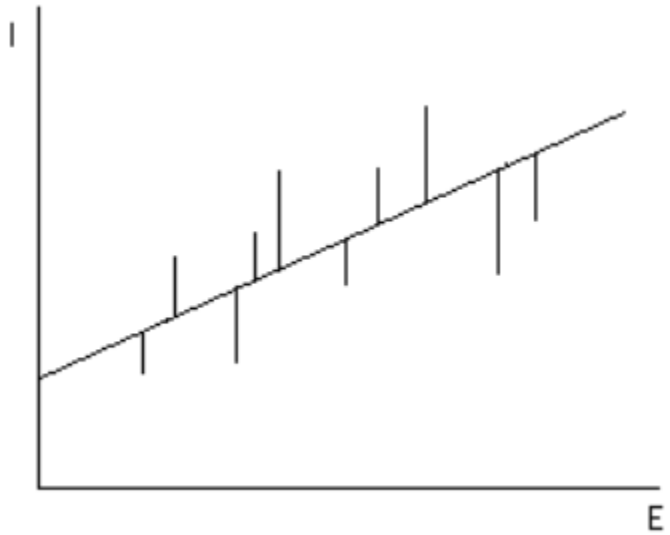
The variable I is termed the “dependent” or “endogenous” variable; E is termed the “independent,” “explanatory,” or “exogenous” variable; α is the “constant term” and β is the “coefficient” of the variable E.

Remember what is observable and what is not. The data set contains observations for I and E. The noise component ε is comprised of unobservable factors, or at least unobserved. The parameters α and β are also unobservable. The task of regression analysis is to produce an estimate of these two parameters, based upon the information contained in the data set and, as shall be seen, upon some assumptions about the characteristics of ε.

To understand how the parameter estimates are generated, note that if we ignore the noise term ε, the equation above for the relationship between I and E is the equation for a line—a line with an “intercept” of α on the vertical axis and a “slope” of β. Returning to the scatter diagram, the hypothesized relationship thus implies that somewhere on the diagram may be found a line with the equation $I = \alpha + \beta E$. The task of estimating α and β is equivalent to the task of estimating where this line is located.

What is the best estimate regarding the location of this line? The answer depends in part upon what we think about the nature of the noise term ε. If we believed that ε was usually a large negative number, for example, we would want to pick a line lying above most or all of our data points—the logic is that if ε is negative, the true value of I (which we observe), given by $I = \alpha + \beta E + \varepsilon$, will be less than the value of I on line $I = \alpha + \beta E$. Likewise, if we believed that ε was systematically positive, a line lying below the majority of data points would be appropriate. Regression analysis assumes, however, that the noise term has no such systematic property, but is on average equal to zero—I will make the assumptions about the noise term more precise in a moment. The assumption that the noise term is usually zero suggests an estimate of the line that lies roughly amid the data, some observations below and some observations above.

But there are many such lines, and it remains to pick one line in particular. Regression analysis does so by embracing a criterion that relates to the estimated noise term or “error” for each observation. To be precise, define the “estimated error” for each observation as the vertical distance between the value of I along with the estimated line $I = \alpha + \beta E$ (generated by plugging the actual value of E into this equation) and the true value of I for the same observation. Superimposing a candidate line on the scatter diagram, the estimated errors for each observation may be seen as follows:



With each possible line that might be superimposed upon the data, a different set of estimated errors will result. Regression analysis then chooses among all possible lines by selecting the one for which the sum of the squares of the estimated errors is at a minimum. This is termed the minimum sum of squared errors (minimum SSE) criterion The intercept of the line chosen by this criterion provides the estimate of α, and its slope provides the estimate of β.

It is hardly obvious why we should choose our line using the minimum SSE criterion. We can readily imagine other criteria that might be utilized (minimizing the sum of errors in absolute value, for example). One virtue of the SSE criterion is that it is very easy to employ computationally. When one expresses the sum of squared errors mathematically and employs calculus techniques to ascertain the values of α and β that minimize it, one obtains expressions for α and β that are easy to evaluate with a computer using only the observed values of E and I in the data sample.[2] But computational convenience is not the only virtue of the minimum SSE criterion—it also has some attractive statistical properties under plausible assumptions about the noise term. These properties will be discussed in a moment after we introduce the concept of multiple regression.

Multiple Regression

Earnings are affected by a variety of factors in addition to years of schooling, factors that were aggregated into the noise term in the simple regression model above. "Multiple regression" is a technique that allows additional factors to enter the analysis separately so that the effect of each can be estimated. It is valuable for quantifying the impact of various simultaneous influences upon a single dependent variable. Further, because of omitted variables bias with simple regression, multiple regression is often essential even when the investigator is only interested in the effects of one of the independent variables.

For purposes of illustration, consider the introduction into the earnings analysis of a second independent variable called "experience." Holding constant the level of education, we would expect someone who has been working for a longer time to earn more. Let X denote years of experience in the labor force and, as in the case of education, we will assume that it has a linear effect upon earnings that is stable across individuals. The modified model may be written:

$$I = \alpha + \beta E + \gamma X + \varepsilon$$

where γ is expected to be positive.

The task of estimating the parameters α, β, and γ is conceptually identical to the earlier task of estimating only α and β. The difference is that we can no longer think of regression as choosing a line in a two-dimensional diagram—with two explanatory variables we need three dimensions, and instead of estimating a line, we are estimating a plane. Multiple regression analysis will select a plane so that the sum of squared errors—the error here is the vertical distance between the actual value of I and the estimated plane—is at a minimum. The intercept of that plane with the I-axis (where E and X are zero) implies the constant term α, its slope in the education dimension implies the coefficient β, and its slope in the experience dimension implies the coefficient γ.

Multiple regression analysis is capable of dealing with an arbitrarily large number of explanatory variables. Though people cannot visualize in more than three dimensions, mathematics does not. With n explanatory variables, multiple regression analysis will estimate the equation of a "hyperplane" in n-space such that the sum of squared errors has been minimized. Its intercept implies the constant term, and its slope in each dimension implies one of the regression coefficients. As in the case of simple regression, the SSE criterion is quite convenient computationally. Formulae for the parameters α, β, γ . . . can be derived readily and evaluated easily on a computer, again using only the observed values of the dependent and independent variables.

The interpretation of the coefficient estimates in multiple regression warrants brief comment. In the model $I = \alpha + \beta E + \gamma X + \varepsilon$, α captures what an individual earns with no education or experience, β captures the effect on the income of a year of education, and γ captures the effect on the income of a year of experience. To put it slightly differently, β is an estimate of the effect of a year of education on income, holding experience constant. Likewise, γ is the estimated effect of a year of experience on income, holding education constant.