Coase-Sandor Working Paper Series in Law and Economics

Coase-Sandor Institute for Law and Economics

1993

An Introduction to Regression Analysis

Alan O. Sykes

Follow this and additional works at: https://chicagounbound.uchicago.edu/law_and_economics

Part of the Law Commons

Recommended Citation

The Inaugural Coase Lecture

An Introduction to Regression Analysis

Alan O. Sykes[*]

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 policy makers as well. Regression has been offered as evidence of liability under Title VII of the Civil Rights Act of 1964,[1] as evidence of racial bias in death penalty litigation,[2] as evidence of damages in contract actions,[3] as evidence of violations under the Voting Rights Act,[4] and as evidence of damages in antitrust litigation,[5] among other things.

In this lecture, I will provide an overview of the most basic techniques of regression analysis—how they work, what they assume,

[1] See, e.g, Bazemore v. Friday, 478 U.S. 385, 400 (1986).

[2] See, e.g., McClesky v. Kemp, 481 U.S. 279 (1987).

[3] See, e.g., Cotton Brothers Baking Co. v. Industrial Risk Insurers, 941 F.2d 380 (5th Cir. 1991).

[4] See, e.g., Thornburgh v. Gingles, 478 U.S. 30 (1986).

[5] See, e.g., Sprayrite Service Corp. v. Monsanto Co., 684 F.2d 1226 (7th Cir. 1982).

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.[6] I wish to emphasize that this lecture is not a comprehensive treatment of the statistical issues that arise in Title VII litigation, and that the discussion of gender discrimination is simply a vehicle for expositing certain aspects of regression technique.[7] 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 single-equation, ordinary least squares regression.

1. 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."

---

[6]See 42 U.S.C. §2000e-2 (1988), as amended.

[7]Readers with a particular interest in the use of regression analysis under Title VII may wish to consult the following references: Campbell, "Regression Analysis in Title VII Cases—Minimum Standards, Comparable Worth, and Other Issues Where Law and Statistics Meet," 36 *Stan. L. Rev.* 1299 (1984); Connolly, "The Use of Multiple Rgeression Analysis in Employment Discrimination Cases," 10 *Population Res. and Pol. Rev.* 117 (1991); Finkelstein, "The Judicial Reception of Multiple Regression Studies in Race and Sex Discrimination Cases," 80 *Colum. L. Rev.* 737 (1980); and Fisher, "Multiple Regression in Legal Proceedings", 80 *Colum. L. Rev.* 702 (1980), at 721–25.

A. *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.
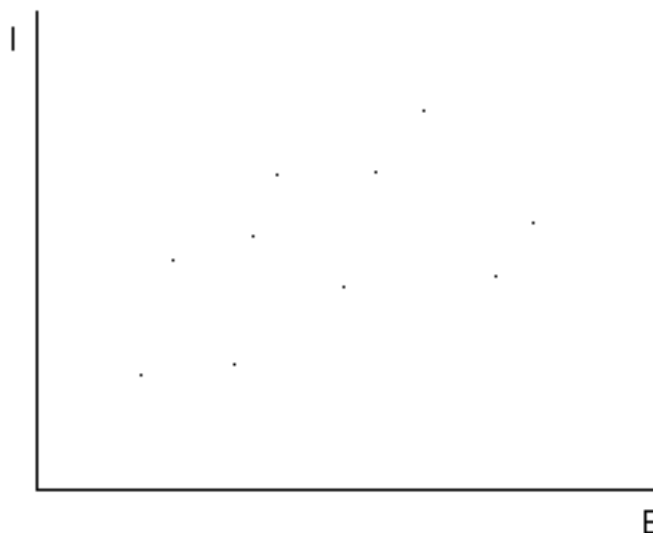
The diagram indeed suggests that higher values of E tend to yield higher values of I, but the relationship is not perfect—it seems that knowledge of E does not suffice for an entirely accurate prediction about I.[8] We can then deduce either that the effect of education upon earnings differs across individuals, or that factors other than education influence earnings. Regression analysis ordinarily embraces the latter explanation.[9] 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

---

[8]More accurately, what one can infer from the diagram is that if knowledge of E suffices to predict I perfectly, then the relationship between them is a complex, nonlinear one. Because we have no reason to suspect that the true relationship between education and earnings is of that form, we are more likely to conclude that knowledge of E is not sufficient to predict I perfectly.

[9]The alternative possibility—that the relationship between two variables is unstable—is termed the problem of "random" or "time varying" coefficients and raises somewhat different statistical problems. See, e.g., H. Theil, *Principles of Econometrics* 622–27 (1971); G. Chow, *Econometrics* 320–47 (1983).

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.[10]

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 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 factors that are unobservable, 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

---

[10]When nonlinear relationships are thought to be present, investigators typically seek to model them in a manner that permits them to be transformed into linear relationships. For example, the relationship $y = cx^{\alpha}$ can be transformed into the linear relationship $\log y = \log c + \alpha \bullet \log x$. The reason for modeling nonlinear relationships in this fashion is that the estimation of linear regressions is much simpler and their statistical properties are better known. Where this approach is infeasible, however, techniques for the estimation of nonlinear regressions have been developed. See, e.g., G. Chow, supra note 9, at 220–51.

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 the 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 in the midst of 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 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:
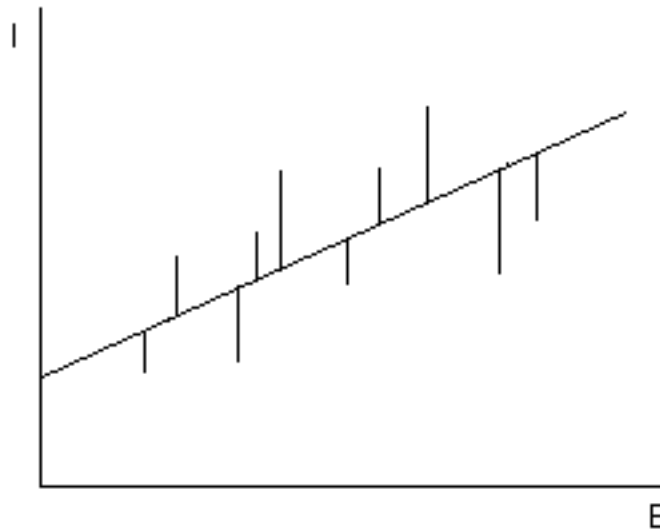
With each possible line that might be superimposed upon the data, a different set of estimated errors will result. Regression analysis then chooses among all possible lines by selecting the one for which the sum of the squares of the estimated errors is at a minimum. This is termed the minimum sum of 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,[11] 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 ob-

---

[11]It should be obvious why simply minimizing the sum of errors is not an attractive criterion—large negative errors and large positive errors would cancel out, so that this sum could be at a minimum even though the line selected fitted the data very poorly.

served values of E and I in the data sample.[12] 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.

B. *Multiple Regression*

Plainly, 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.

---

[12]The derivation is so simple in the case of one explanatory variable that it is worth including here: Continuing with the example in the text, we imagine that we have data on education and earnings for a number of individuals, let them be indexed by j. The actual value of earnings for the jth individual is I_j and its estimated value for any line with intercept α and slope β will be $\alpha + \beta E_j$. The estimated error is thus $I_j - \alpha - \beta E_j$. The sum of squared errors is then $\sum_j (I_j - \alpha - \beta E_j)^2$. Minimizing this sum with respect to α requires that its derivative with respect to α be set to zero, or $-2\sum_j (I_j - \alpha - \beta E_j) = 0$. Minimizing with respect to β likewise requires $-2\sum_j E_j (I_j - \alpha - \beta E_j) = 0$. We now have two equations in two unknowns that can be solved for α and β.

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 being 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 in fact capable of dealing with an arbitrarily large number of explanatory variables. Though people lack the capacity to 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.[13]

The interpretation of the coefficient estimates in a 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 income of a year of education, and γ captures the effect on 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,

---

[13]The derivation may be found in any standard econometrics text. See, e.g., E. Hanushek and J. Jackson, *Statistical Methods for Social Scientists* 110–16 (1977); J. Johnston, *Econometric Methods* 122–32 (2d ed. 1972).

holding experience constant. Likewise, γ is the estimated effect of a year of experience on income, holding education constant.

2. Essential Assumptions and Statistical Properties of Regression

As noted, the use of the minimum SSE criterion may be defended on two grounds: its computational convenience, and its desirable statistical properties. We now consider these properties and the assumptions that are necessary to ensure them.[14]

Continuing with our illustration, the hypothesis is that earnings in the "real world" are determined in accordance with the equation $I = \alpha + \beta E + \gamma X + \varepsilon$—true values of α, β, and γ exist, and we desire to ascertain what they are. Because of the noise term ε, however, we can only estimate these parameters.

We can think of the noise term ε as a random variable, drawn by nature from some probability distribution—people obtain an education and accumulate work experience, then nature generates a random number for each individual, called ε, which increases or decreases income accordingly. Once we think of the noise term as a random variable, it becomes clear that the *estimates* of α, β, and γ (as distinguished from their true values) will also be random variables, because the estimates generated by the SSE criterion will depend upon the particular value of ε drawn by nature for each individual in the data set. Likewise, because there exists a probability distribution from which each ε is drawn, there must also exist a probability distribution from which each parameter estimate is drawn, the latter distribution a function of the former distributions. The attractive statistical properties of regression all concern the relationship between the probability distribution of the parameter estimates and the true values of those parameters.

We begin with some definitions. The minimum SSE criterion is termed an *estimator.* Alternative criteria for generating parameter estimates (such as minimizing the sum of errors in absolute value) are also estimators.

---

[14]An accessible and more extensive discussion of the key assumptions of regression may be found in Fisher, supra note 7.

Each parameter estimate that an estimator produces, as noted, can be viewed as a random variable drawn from some probability distribution. If the mean of that probability distribution is equal to the true value of the parameter that we are trying to estimate, then the estimator is *unbiased.* In other words, to return to our illustration, imagine creating a sequence of data sets each containing the same individuals with the same values of education and experience, differing only in that nature draws a different ε for each individual for each data set. Imagine further that we recompute our parameter estimates for each data set, thus generating a range of estimates for each parameter α, β, and γ. If the estimator is unbiased, we would find that on average we recovered the true value of each parameter.

An estimator is termed *consistent* if it takes advantage of additional data to generate more accurate estimates. More precisely, a consistent estimator yields estimates that converge on the true value of the underlying parameter as the sample size gets larger and larger. Thus, the probability distribution of the estimate for any parameter has lower variance[15] as the sample size increases, and in the limit (infinite sample size) the estimate will equal the true value.

The variance of an estimator for a *given* sample size is also of interest. In particular, let us restrict attention to estimators that are unbiased. Then, lower variance in the probability distribution of the estimator is clearly desirable[16]—it reduces the probability of an estimate that differs greatly from the true value of the underlying parameter. In comparing different unbiased estimators, the one with the lowest variance is termed *efficient* or *best.*

Under certain assumptions, the minimum SSE criterion has the characteristics of unbiasedness, consistency, and efficiency—these assumptions and their consequences follow:

---

[15]"Variance" is a mesaure of the dispersion of the probability distribution of a random variable. Consider two random variables with the same mean (same average value). If one of them has a distribution with greater variance, then, roughly speaking, the probability that the variable will take on a value far from the mean is greater.

[16]Lower variance by itself is not necessarily an attractive property for an estimator. For example, we could employ an estimator for β of the form "$\beta = 17$" irrespective of the information in the data set. This estimator has zero variance.

(1) If the noise term for each observation, ε, is drawn from a distribution that has a mean of zero, then the sum of squared errors criterion generates estimates that are unbiased and consistent.

That is, we can imagine that for each observation in the sample, nature draws a noise term from a different probability distribution. As long as each of these distributions has a mean of zero (even if the dsitributions are not the same), the minimum SSE criterion is unbiased and consistent.[17] This assumption is logically sufficient to ensure that one other condition holds—namely, that each of the explanatory variables in the model is uncorrelated with the expected value of the noise term.[18] This will prove important later.

(2) If the distributions from which the noise terms are drawn for each observation have the same variance, and the noise terms are statistically independent of each other (so that if there is a positive noise term for one observation, for example, there is no reason to expect a positive or negative noise term for any other observation), then the sum of squared errors criterion gives us the best or most efficient estimates available from any *linear* estimator (defined as an estimator that computes the parameter estimates as a linear function of the noise term, which the SSE criterion does).[19]

If assumptions (2) are violated, the SSE criterion remains unbiased and consistent but it is possible to reduce the variance of the estimator by taking account of what we know about the noise term. For example, if we know that the variance of the distribution from which the noise term is drawn is bigger for certain observations, then the size of the noise term for those observations is *likely* to be larger. And, because the noise is larger, we will want to give those observations less weight in our analysis. The statistical procedure for dealing

---

[17]See, e.g, P. Kennedy, *A Guide to Econometrics* 42–44 (2d ed. 1985).

[18]If the expected value of the noise term is always zero irrespective of the values of the explanatory variables for the observation with which the noise term is associated, then by definition the noise term cannot be correlated with any explanatory variable.

[19]E.g., id. at 44; J. Johnston, supra note 13, at 126–27.

with this sort of problem is termed "generalized least squares," which is beyond the scope of this lecture.[20]

3. An Illustration—Discrimination on the Basis of Gender

To illustrate the ideas to this point, as well as to suggest how regression may have useful applications in a legal proceeding, imagine a hypothetical firm that has been sued for wage discrimination on the basis of gender. To investigate these allegations, data have been gathered for all of the firm's employees. The questions to be answered are (a) whether discrimination is occurring (liability), and (b) what its consequences are (damages). We will address them using a modified version of the earnings model developed in section 1.

The usefulness of *multiple* regression here should be intuitively apparent. Suppose, for example, that according to the data, women at the firm on average make less than men. Is this fact sufficient to establish actionable discrimination? The answer is no if the difference arises because women at this firm are less well-educated, for example (and thus by inference less productive), or because they are less experienced.[21] In short, the legal question is whether women earn less after all of the factors that the firm may permissibly consider in setting wages have been taken into account.

To generate the data for this illustration, I assume a hypothetical "real world" in which earnings are determined by equation (1):

$$\text{Earnings} = 5000 + 1000 \cdot \text{School} + 50 \cdot \text{Aptitude} + 300 \cdot \text{Experience} - 2000 \cdot \text{Gendum} + \text{Noise} \qquad (1)$$

where "School" is years of schooling; "Aptitude" is a score between 100 and 240 on an aptitude test; "Experience" is years of experience in the work force; and "Gendum" is a variable that equals 1 for women and zero for men (more about this variable in a moment). To produce the artificial data set, I made up fifty observations (corresponding to fifty fictitious individuals) for each of the explanatory variables, half men and half women. In making up the data, I deliberately tried to introduce some positive correlation between the

---

[20]See, e.g., id. at 208–66.

[21]See, e.g., Miller v Kansas Electric Power Cooperative, Inc., 1990 WL 120935 (D. Kan.).

schooling and aptitude variables, for reasons that will become clear later. I then employed a random number generator to produce a noise term drawn from a normal distribution, with a standard deviation (the square root of the variance) equal to 3,000 and a mean of zero. This standard deviation was chosen more or less arbitrarily to introduce a considerable but not overwhelming amount of noise in proportion to the total variation in earnings. The right-hand-side variables were then used to generate the "actual value" of earnings for each of the fifty "individuals."

The effect of gender on earnings in this hypothetical firm enters through the variable Gendum. Gendum is a "dummy" variable in econometric parlance because its numerical value is arbitrary, and it simply captures some nonnumerical attribute of the sample population. By construction here, men and women both earn the same returns to education, experience, and aptitude, but holding these factors constant the earnings of women are $2,000 lower. In effect, the constant term (baseline earnings) is lower for women, but otherwise women are treated equally. In reality, of course, gender discrimination could arise in other ways (such as lower returns to education and experience for women, for example), and I assume that it takes this form only for purposes of illustration.

Note that the random number generator that I employed here generates noise terms with an expected value of zero, each drawn from a distribution with the same variance. Further, the noise terms for the various observations are statistically independent (the realized value of the noise term for each observation has no influence on the noise term drawn for any other observation). Hence, the noise terms satisfy the assumptions necessary to ensure that the minimum SSE criterion yields unbiased, consistent, and efficient estimates. The expected value of the estimate for each parameter is equal to the true value, therefore, and no other linear estimator will do a better job at recovering the true parameters than the minimum SSE criterion. It is nevertheless interesting to see just how well regression analysis performs. I used a standard computer package to estimate the constant term and the coefficients of the four independent variables from the "observed" values of Earnings, School, Aptitude, Experience, and Gendum for each of the fifty hypothetical individuals. The results are reproduced in table 1, under the column labeled

"Estimated Value." (We will discuss the last three columns and the R^2 statistic in the next section.)

Table 1 (noise term with standard deviation of 3,000)

Variable	"True value"	Estimated value	Standard error	t-statistic	Prob (2-tail)
Constant	5000.0	4136.7	3781.8	1.094	.280
School	1000.0	1584.6	288.1	5.500	.000
Aptitude	50.0	6.4	27.3	0.236	.814
Experience	300.0	241.7	80.8	2.992	.004
Gendum	−2000.0	−1470.4	1402.2	−1.049	.300

$R^2 = .646$

Note that all of the estimated parameters have the right sign. Just by chance, it turns out that the regression overestimates the returns to schooling and underestimates the other parameters. The estimated coefficient for Aptitude is off by a great deal in proportion to its true value, and in a later section I will offer an hypothesis as to what the problem is. The other parameter estimates, though obviously different from the true value of the underlying parameter, are much closer to the mark. With particular reference to the coefficient of Gendum, the regression results correctly suggest the presence of gender discrimination, though its magnitude is underestimated by about 25 percent (remember that an overestimate of the same magnitude was just as likely ex ante, that is, before the actual values for the noise terms were generated).

The source of the error in the coefficient estimates, of course, is the presence of noise. If the noise term were equal to zero for every observation, the true values of the underlying parameters could be recovered in this illustration with perfect accuracy from the data for only five hypothetical individuals—it would be a simple matter of solving five equations in five unknowns. And, if noise is the source of error in the parameter estimates, intuition suggests that the magnitude of the noise will affect the accuracy of the regression estimates, with more noise leading to less accuracy on average. We will make this intuition precise in the next section, but before

proceeding it is perhaps useful to repeat the parameter estimation experiment for a hypothetical firm in which the data contain less noise. To do so, I took the "data" for the independent variables used in the experiment above and again generated values for earnings for the fifty hypothetical individuals using equation (1), changing only the noise terms. This time, the noise terms were drawn by the random number generator from a normal distribution with standard deviation of 1,000 rather than 3,000 (a significant reduction in the amount of noise). Reestimating the regression parameters from this modified data set produced the results in table 2:

Table 2 (noise term with standard deviation of 1,000)

Variable	"True value"	Estimated value	Standard error	t-statistic	Prob (2-tail)
Constant	5000.0	4784.2	945.4	5.060	.000
School	1000.0	1146.2	72.0	15.913	.000
Aptitude	50.0	39.1	6.8	5.741	.000
Experience	300.0	285.4	20.2	14.131	.000
Gendum	−2000.0	−1867.6	350.5	−5.328	.000

$R^2 = .964$

Not surprisingly, the estimated parameters here are considerably closer to their true values. It was not certain that they would be, because after all their expected values are equal to their true values regardless of the amount of noise (the estimator is unbiased). But on average we would expect greater accuracy, and greater accuracy indeed emerges here. Put more formally, the probability distributions of the parameter estimates have greater variance, the greater the variance of the noise term. The variance of the noise term thus affects the degree of confidence that we have in the accuracy of regression estimates.

In real applications, of course, the noise term is unobservable, as is the distribution from which it is drawn. The variance of the noise term is thus unknown. It can, however, be estimated using the difference between the predicted values of the dependent variable for each observation and the actual value (the "estimated errors" defined

earlier). This estimate in turn allows the investigator to assess the explanatory power of the regression analysis and the "statistical significance" of its parameter estimates.

4. Statistical Inference and Goodness of fit

Recall that the parameter estimates are themselves random variables, dependent upon the random variables ε. Thus, each estimate can be thought of as a draw from some underlying probability distribution, the nature of that distribution as yet unspecified. With a further assumption, however, we can compute the probability distribution of the estimates, and use it to test hypotheses about them.

A. Statistical Inference

Most readers are familiar, at least in passing, with a probability distribution called the "normal." Its shape is that of a "bell curve," indicating among other things that if a sample is drawn from the distribution, the most likely values for the observations in the sample are those close to the mean and least likely values are those farthest from the mean. *If we assume that the noise terms ε are all drawn from the same normal distribution,* it is possible to show that the parameter estimates have a normal distribution as well.[22]

The variance of this normal distribution, however, depends upon the variance of the distribution from which the noise terms are drawn. This variance is unknown in practice and can only be estimated using the estimated errors from the regression to obtain an estimate of the variance of the noise term. The estimated variance of the noise term in turn can be used to construct an estimate of the

---

[22]See, e.g., E. Hanushek and J. Jackson, supra note 13, at 66–68; J. Johnston, supra note 13, at 135–38. The supposition that the noise terms are normally distributed is often intuitively plausible and may be loosely justified by appeal to "central limit theorems," which hold that the average of a large number of random variables tends toward a normal distribution even if the individual random variables that enter into the average are not normally distributed. See, e.g., R. Hogg and A. Craig, *Introduction to Mathematical Statistics* 192–95 (4th ed. 1978); W. Feller, *An Introduction to Probability Theory and Its Applications,* vol. 1, 243–48 (3d ed. 1968). Thus, if we think of the noise term as the sum of a large number of independent, small disturbances, theory affords considerable basis for the supposition that its distribution is approximately normal.

variance of the normal distribution for each coefficient. The square root of this estimate is called the "standard error" of the coefficient—call this measure *s.*

It is also possible to show[23] that if the parameter estimate, call it π, is normally distributed with a mean of μ, then $(\pi - \mu)/s$ has a "Student's t" distribution. The t-distribution looks very much like the normal, only it has "fatter" tails and its mean is zero. Using this result, suppose we hypothesize that the true value of a parameter in our regression model is μ. Call this the "null hypothesis." Because the minimum SSE criterion is an unbiased estimator, we can deduce that our parameter estimate is drawn from a normal distribution with a mean of μ if the null hypothesis is true. If we then subtract μ from our actual parameter estimate and divide by its standard error, we obtain a number called the *t-statistic,* which is drawn from a t-distribution if the null hypothesis is true. This statistic can be positive or negative as the parameter estimate from which it is derived is greater or less than the hypothesized true value of the parameter. Recalling that the t-distribution is much like a normal with mean of zero, we know that large values of the t-statistic (in absolute value) will be drawn considerably less frequently than small values of the t-statistic. And, from the construction of the t-statistic, large values for that statistic arise (in absolute value), other things being equal, when the parameter estimate on which it is based differs from its true (hypothesized) value by a great deal.

This insight is turned on its head for hypothesis testing. We have just argued that a large t-statistic (in absolute value) will arise fairly infrequently if the null hypothesis is correct. Hence, when a large t-statistic *does* arise, it will be tempting to conclude that the null hypothesis is false. The essence of hypothesis testing with a regression coefficient, then, is to formulate a null hypothesis as to its true value, and then to decide whether to accept or reject it according to whether the t-statistic associated with that null hypothesis is large enough that the plausibility of the null hypothesis is sufficiently in doubt.[24]

---

[23]See sources cited note 22 supra.

[24]I limit the discussion here to hypothesis testing regarding the value of a particular parameter. In fact, other sorts of hypotheses may readily be tested, such

One can be somewhat more precise. We might resolve that the null hypothesis is implausible if the t-statistic associated with our regression estimate lies so far out in one tail of its t-distribution that such a value, or one even larger in absolute value, would arise less than, say, 5 percent of the time if the null hypothesis is correct. Put differently, we will reject the null hypothesis if the t-statistic falls either in the uppermost tail of the t-distribution, containing 2.5 percent of the draws representing the largest positive values, or in the lowermost tail, containing 2.5 percent of the draws representing the largest negative values. This is called a "two-tailed test."

Alternatively, we might have a strong prior belief about the true value of a parameter that would lead us to accept the null hypothesis even if the t-statistic lies far out in *one* of the tails of the distribution. Consider the coefficient of the gender dummy in table 1 as an illustration. Suppose the null hypothesis is that the true value of this coefficient is zero. Under what circumstances would we reject it? We might find it implausible that the true value of the coefficient would be positive, reflecting discrimination *against* men. Then, even if the estimated coefficient for the gender dummy is positive with a large positive t-statistic, we would still accept the null hypothesis that its true value is zero. Only a negative coefficient estimate with a large negative t-statistic would lead us to conclude that the null hypothesis was false. Where we reject the null hypothesis only if a t-statistic that is large in absolute value has a particular sign, we are employing a "one-tailed test."

To operationalize either a one- or two-tailed test, it is necessary to compute the exact probability of a t-statistic as large or larger in absolute value as the one associated with the parameter estimate at issue. In turn, it is necessary to know exactly how "spread out" is the t-distribution from which the estimate has been drawn. A further parameter that we need to pin down the shape of the t-distribution in this respect is called the "degrees of freedom," defined as the number of observations in the sample less the number of parameters to be estimated. In the illustrations of tables 1 and 2, we have fifty observations in the sample, and we are estimating 5 parameters, so

---

as the hypothesis that all parameters in the model are zero, the hypothesis that some subset of the parameters are zero, and so on.

the t-distribution for any of the parameter estimates has 45 degrees
of freedom. The fewer the degrees of freedom, the more "spread
out" is the t-distribution and thus the greater is the probability of
drawing large t-statistics. The intuition is that the larger the sample,
the more collapsed is the distribution of any parameter estimate
(recall the concept of consistency above). By contrast, the more
parameters we seek to estimate from a sample of given size, the more
information we are trying to extract from the data and the less
confident we can be in the estimate of each parameter—hence, the
associated t-distribution is more "spread out."[25]

Knowing the degrees of freedom for the t-distribution allows an
investigator to compute the probability of drawing the t-statistic in
question, or one larger in absolute value, assuming the truth of the
null hypothesis. Using the appropriate one- or two-tailed test (the
former necessary only when the t-statistic is of the right sign), the
investigator then rejects the null hypothesis if this probability is
sufficiently small.

But what do we mean by "sufficiently small"? The answer is by
no means obvious, and it depends upon the circumstances. It has be-
come convention in social scientific research to test one particular
null hypothesis—namely, the hypothesis that the true value of a co-
efficient is zero. Under this hypothesis, μ in our notation above is
equal to zero, and hence the t-statistic is simply π/s, the coefficient
estimate divided by its standard error. It is also convention to em-
brace a "significance level" of .10, .05, or .01—that is, to inquire
whether the t-statistic that the investigator has obtained, or one even
larger in absolute value, would arise more than 10 percent, 5 percent,
or 1 percent of the time when the null hypothesis is correct. Where
the answer to this question is no, the null hypothesis is rejected and
the coefficient in question is said to be "statistically significant." For
example, if the parameter estimate that was obtained is far enough
from zero that an estimate of that magnitude, or one even further
from zero, would arise less than 5 percent of the time, then the co-
efficient is said to be significant at the .05 level.

The question whether the conventional social scientific signifi-
cance tests are appropriate when regression analysis is used for legal

---

[25]See sources cited note 13 supra.

applications, particularly in litigation, is a difficult one that I will defer to the concluding section of this lecture. I will simply assume for now that we are interested in the general problem of testing some null hypothesis, and that we will reject it if the parameter estimate obtained lies far enough out in one of the tails of the distribution from which the estimate has been drawn. We leave open the question of what constitutes "far enough" and simply seek to compute the probability under a one- or two-tailed test of obtaining an estimate as far from the mean of the distribution as that generated by the regression if the null hypothesis is true.

Most computerized regression packages report not only the parameter estimate itself (π in our notation), but also the standard error of each parameter estimate (s in our notation). This value, coupled with the hypothesized true parameter value (μ in our notation), can then be employed to generate the appropriate t-statistic for any null hypothesis. Many regression packages also report a number called the "t-statistic," which is invariably based upon the conventional social scientific null hypothesis that the true parameter value is zero. finally, some packages report the probability that the t-statistic at issue could have been generated from a t-distribution with the appropriate degrees of freedom under a one- or two-tailed test.[26]

Returning to tables 1 and 2, all of this information is reported for each of the five parameter estimates—the standard error, the value of the t-statistic for the null hypothesis that the true parameter value is zero, and the probability of getting a t-statistic that large or larger in absolute value under a two-tailed test with 45 degrees of freedom. To interpret this information, consider the estimated coefficient for the

---

[26]If the regression package does not report these probabilities, they can readily be found elsewhere. It has become common practice to include in statistics and econometrics books tables of probabilities for t-distributions with varying degrees of freedom. Knowing the degrees of freedom associated with a t-statistic, therefore, one can consult such a table to ascertain the probability of obtaining a t-statistic as far from zero or farther as the one generated by the regression (the concept "far from zero" again defined by either a one- or two-tailed test). As a point of reference, when the degrees of freedom are large (say, 50 or more), then the .05 significance level for a two-tailed test requires a t-statistic approximately equal to 2.0.

gender dummy in table 1. The estimated coefficient of –1470.4 has standard error of 1402.2 and thus a t-statistic of –1470.4/1402.2 = –1.049. The associated probability under a two-tailed test is reported as .30. This means that if the true value of the coefficient for the gender dummy were zero, a coefficient greater than or equal to 1470.4 in absolute value would nevertheless arise 30 percent of the time given the degrees of freedom of the t-distribution from which the coefficient estimate is drawn. A rejection of the null hypothesis on the basis of a parameter estimate equal to 1470.4 or greater in absolute value, therefore, will be erroneous three times out of ten when the null hypothesis is true. By conventional social science standards, therefore, the significance level here is too low to reject the null hypothesis, and the coefficient of the gender dummy is not statistically significant. It is noteworthy that in this instance (in contrast to any real world application), we know the true parameter value, namely –2000.0. Hence, if we employ a conventional two-tailed significance test, we are led erroneously to reject the hypothesis that gender discrimination is present.

As noted, we may regard the two-tailed test as inappropriate for the coefficient of the gender dummy because we find the possibility of discrimination against men to be implausible. It is a simple matter to construct an alternative one-tailed test: Table 1 indicates that a coefficient estimate of 1470.4 or greater in absolute value will occur 30 percent of the time if the true value of the coefficient is zero. Put differently, an estimate of the gender dummy coefficient greater than or equal to 1470.4 will arise 15 percent of the time, and an estimate less than or equal to –1470.4 will arise 15 percent of the time. It follows that if we are only interested in the lower tail of the t-distribution, rejection of the null hypothesis (when it is true) will be erroneous only 15 percent of the time if we require a parameter estimate of –1470.4 or smaller. The one-tailed significance level is thus .15, still below the conventional thresholds for statistical significance.[27] Using such significance levels, therefore, we again are led to accept the null hypothesis, in this case erroneously.

---

[27]The result in this illustration is general—for any t-statistic, the probability of rejecting the null hypothesis erroneously under a one-tailed test will be exactly half that probability under a two-tailed test.

I offer this illustration not to suggest that there is anything wrong with conventional significance tests, but simply to indicate how one reduces the chance of erroneously rejecting the null hypothesis (call this a "Type I" error) only by increasing the chance of erroneously accepting it (call this a "Type II" error). The conventional significance tests implicitly give great weight to the importance of avoiding Type I errors, and less weight to the avoidance of Type II errors, by requiring a high degree of confidence in the falsity of the null hypothesis before rejecting it. This seems perfectly appropriate for most scientific applications, in which the researcher is justifiably asked to bear a considerable burden of proof before the scientific community will accept that the data establish an asserted causal relation. Whether the proponent of regression evidence in a legal proceeding should bear that same burden of proof is a more subtle issue.

B. *Goodness of fit*

Another common statistic associated with regression analysis is the R^2. This has a simple definition—it is equal to one minus the ratio of the sum of squared *estimated* errors (the deviation of the actual value of the dependent variable from the regression line) to the sum of squared deviations about the mean of the dependent variable. Intuitively, the sum of squared deviations about its mean is a measure of the total variation of the dependent variable. The sum of squared deviations about the regression line is a measure of the extent to which the regression fails to explain the dependent variable (a measure of the noise). Hence, the R^2 statistic is a measure of the extent to which the total variation of the dependent variable *is* explained by the regression. It is not difficult to show that the R^2 statistic necessarily takes on a value between zero and one.[28]

A high value of R^2, suggesting that the regression model explains the variation in the dependent variable well, is obviously important if one wishes to use the model for predictive or forecasting purposes. It is considerably less important if one is simply interested in particular parameter estimates (as, for example, if one is searching for evidence of discrimination, as in our illustration, and thus

---

[28]See, e.g., E. Hanushek and J. Jackson, supra note 13, at 57–58.

focused on the coefficient of the gender dummy). To be sure, a large unexplained variation in the dependent variable will increase the standard error of the coefficients in the model (which are a function of the estimated variance of the noise term), and hence regressions with low values of R^2 will often (but by no means always) yield parameter estimates with small t-statistics for any null hypothesis. Because this consequence of a low R^2 will be reflected in the t-statistics, however, it does not afford any reason to be concerned about a low R^2 per se.

As a quick illustration, turn back to tables 1 and 2. Recall that the noise terms for the data set from which the estimates in table 1 were generated were drawn from a distribution with a standard deviation of 3,000, while for table 2 the noise terms were drawn from a distribution with a standard deviation of 1,000. The unexplained variation in the earnings variable is likely to be greater in the first data set, therefore, and indeed the R^2 statistics confirm that it is (.646 for table 1 and .964 for table 2). Likewise, because the estimated variance of the noise term is greater for the estimates in table 1, we expect the coefficient estimates to have larger standard errors and smaller t-statistics. This expectation is also borne out on inspection of the two tables. Variables with coefficients that are statistically significant by conventional tests in table 2, therefore, such as the gender dummy, are not statistically significant in table 1.

In these illustrations, the value of R^2 simply reflects the amount of noise in the data, and a low R^2 is not inconsistent with the minimum SSE criterion serving as an unbiased, consistent, and efficient estimator because we know that the noise terms were all independent draws from the same distribution with a zero mean. In practice, however, a low value of R^2 *may* indicate that important and systematic factors have been omitted from the regression model. This possibility raises again the concern about omitted variables bias.

5. *Two Common Statistical Problems in Regression Analysis*

Much of the typical econometrics course is devoted to what happens when the assumptions that are necessary to make the minimum SSE criterion unbiased, consistent, and efficient do not hold. I cannot begin to provide a full sense of these issues in such a brief lecture

and will simply illustrate two of the many complications that may arise, chosen because they are both common and quite important.

ᴀ. *Omitted Variables*

As noted, the omission from a regression of some variables that affect the dependent variable may cause an "omitted variables bias." The problem arises because any omitted variable becomes part of the noise term, and the result may be a violation of the assumption necessary for the minimum SSE criterion to be an unbiased estimator.

Recall that assumption—that each noise term is drawn from a distribution with a mean of zero. We noted that this assumption logically implies the absence of correlation between the explanatory variables included in the regression and the expected value of the noise term (because whatever the value of any explanatory variable, the expected value of the noise term is always zero). Thus, suppose we start with a properly specified model in which the noise term for every observation has an expected value of zero. Now, omit one of the independent variables. If the effect of this variable upon the dependent variable is not zero for each observation, the new noise terms now come from distributions with nonzero means. One consequence is that the estimate of the constant term will be biased (part of the estimated value for the constant term is actually the mean effect of the omitted variable). Further, unless the omitted variable is uncorrelated with the included ones, the coefficients of the included ones will be biased because they now reflect not only an estimate of the effect of the variable with which they are associated, but also partly the effects of the omitted variable.[29]

To illustrate the omitted variables problem, I took the data on which the estimates reported in table 1 are based, and reran the regression after omitting the schooling variable. The results are shown in table 3:

---

[29]See J. Johnston, supra note 13, at 168–69; E. Hanushek and J. Jackson, supra note 13, at 81–82. The bias is a function of two things—the true coefficients of the excluded variables, and the correlation within the data set between the included and the excluded variables.

Table 3. Omitted variable illustration

Variable	"True value"	Estimated value	Standard error	t-statis-tic	Prob (2-tail)
Constant	5000.0	9806.5	4653.8	2.107	.041
School	1000.0	omitted			
Aptitude	50.0	107.5	25.6	4.173	.000
Experience	300.0	256.9	103.3	2.487	.017
Gendum	−2000.0	−2445.5	1779.0	−1.375	.176

R^2 = .408

You will note that the omission of the schooling variable lowers the R^2 of the regression, which is not surprising given the original importance of the variable. It also alters the coefficient estimates. The estimate for the constant term rises considerably, because the mean effect of schooling on income is positive. It is not surprising that the constant term is thus estimated to be greater than its true value. An even more significant effect of the omission of schooling is on the coefficient estimate for the aptitude variable, which increases dramatically from below its true value to well above its true value and becomes highly significant. The reason is that the schooling variable is highly correlated (positively) with aptitude in the data set—the correlation is .69—and because schooling has a positive effect on earnings. Hence, with the schooling variable omitted, the aptitude coefficient is erroneously capturing some of the (positive) returns to education as well as the returns to "aptitude." The consequence is that the minimum SSE criterion yields an upward biased estimate of the coefficient for aptitude, and in this case the actual estimate is indeed above the true value of that coefficient.

The effect on the other coefficients is more modest, though non-trivial. Notice, for example, that the coefficient of gendum increases (in absolute value) significantly. This is because schooling happens to be positively correlated with being male in my fictitious data set—without controlling for schooling, the apparent effect of gender is exaggerated because females are somewhat less well educated on average.

The omitted variables problem is troublesome to investigators not simply because it requires them to collect data on more variables to avoid it, but because the omitted variables are often unobservable. Real world studies of gender discrimination are perhaps a case in point. One can readily imagine that earnings depend on such factors as innate ability and motivation, both of which may be unobservable to an investigator. Omitted variables bias may then become something that the investigator cannot avoid, and an understanding of its consequences then becomes doubly important. For an investigator concerned primarily with the coefficient of the gender dummy, it might be argued, the omitted variables bias caused by the exclusion of innate ability and motivation should be modest because the correlation in the sample between gender and those omitted variables might plausibly be assumed to be small. Where the problem appears likely to be serious, by contrast, the utility of conventional regression as an investigative tool diminishes considerably.[30]

I note in passing that the problem of including extraneous or irrelevant variables is less serious. Their expected coefficient is zero and the estimates of the other coefficients are not biased, although the efficiency of the minimum SSE criterion is lessened.[31]

I also note in passing a problem that is closely related to the omitted variables problem, termed "errors in variables." In many regression studies, it is inevitable that some explanatory variables will be measured with error. Such errors become part of the noise term. Let us assume that in the absence of measurement error, the noise terms obey the assumption needed for unbiasedness and consistency—they are all drawn from a distribution with zero mean and are thus uncorrelated with the explanatory variables. With measurement error, however, this assumption will no longer hold.

---

[30]Econometricians have developed some more sophisticated regression techniques to deal with the problem of unobservable variables, but these are not always satisfactory because of certain restrictive assumptions than an investigator must make in using them. See, e.g., Griliches, "Errors in Variables and Other Unobservables," 42 *Econometrica* 971 (1974). An accessible discussion of the omitted variables problem and related issues may be found in P. Kennedy, supra note 17, at 69–72.

[31]Id.

Imagine, for concreteness, that earnings depend on education, experience, and so on, as hypothesized earlier, and on innate ability as suggested above. Instead of supposing that innate ability is an omitted variable, however, suppose that the aptitude test score included in the regression is a "proxy" for innate ability. That is, we regard it as an imperfect measure of ability, correlated with it but not perfectly. When the test score underestimates ability, the noise term rises, and when the aptitude score overestimates ability the noise term falls. The result is a negative correlation between the noise term and the aptitude/ability variable. Put differently, if the noise term without measurement error is drawn from a distribution with zero mean, then the noise term including the measurement error is drawn from a distribution with a mean equal to the magnitude of that error. The consequence once again is bias in the estimated coefficients of the model.[32]

B. *Multicollinearity*

The multicollinearity problem does not result in biased coefficient estimates, but does increase the standard error of the estimates and thus reduces the degree of confidence that one can place in them. The difficulty arises when two independent variables are closely correlated, creating a situation in which their effects are difficult to separate.

The following illustration will convey the essential intuition: Suppose that two law school faculty members (call them Baird and Picker) regularly address alumni luncheons, held partly for the purpose of stimulating alumni contributions. Assume that each time one gives a luncheon speech, the other does too, and that the only available datum on alumni contributions is aggregate monthly

---

[32]One standard technique for addressing this problem is termed "instrumental variables," which replaces the tainted variable with another variable that is thought to be closely associated with it but also thought uncorrelated with the disturbance term. For a variety of reasons, however, the instrumental variables technique is not satisfactory in many cases, and the errors in variables problem is consequently one of the most serious difficulties in the use of regression techniques. A discussion of the instrumental variables technique and other possible responses to the errors in variables problem may be found in P. Kennedy, supra note 17, at 113–16; J. Johnston, supra note 13, at 281–91.

giving. We somehow know that each time both give a speech in a month, alumni contributions rise by $10,000. When they each give two speeches in a month, contributions rise by $20,000, and so on.

Thus, by hypothesis, we know the *joint* effect of a speech by both Baird and Picker ($10,000), but nothing in the data permit us to ascertain their individual effects. Perhaps each speech increases contributions by $5,000, but it might be that one speaker induces an extra $10,000 in giving and the other none, or that one speaker induces an extra $30,000 in giving and the other reduces giving by $20,000. In econometric parlance, the data on speeches given by Baird and Picker are perfectly collinear—the correlation between the number of speeches per month by Baird and the number per month by Picker is 1.0. An attempt to estimate the effect of the number of speeches given by each upon contributions would fail and result in an error message from the computer (for reasons that we need not detail, it would find itself trying to divide by zero).

The term "multicollinearity" usually refers to a problem in the data short of the perfect collinearity in our illustration, but where changes in two variables are nevertheless highly correlated to the point that it is difficult to separate their effects. Because multicollinearity does not go to any property of the noise term, the minimum SSE criterion can still be unbiased, consistent, and efficient. But the difficulty in separating the effects of the two variables introduces greater uncertainty into the estimator, manifest as an increase in the standard errors of the coefficients and a reduction in their t-statistics.

One illustration of the effects of multicollinearity may already have been provided. In our discussion of table 1, we noted that the coefficient estimate for the aptitude variable was far below its true value. As it turns out, aptitude and schooling are highly correlated in the data set, and this affords a plausible conjecture as to why the coefficient for the schooling variable is too high and that for aptitude insignificantly small (some of the effects of aptitude in the sample are captured by the schooling coefficient).

To give another illustration, which incidentally allows us to introduce another use of "dummy" variables, suppose that gender discrimination at our hypothetical firm affects the earnings of women in two ways—through an effect on the baseline earnings of women as

before, and through an effect on the returns to education for women. In particular, recall that in equation (1) both sexes earned $1,000 per year of schooling. Suppose now that males earn $1,000, but females earn only $800. This effect can be captured mathematically by an "interaction term" incorporating the gender dummy, so that earnings are now determined in accordance with equation (2):

$$\text{Earnings} = 5000 + 1000 \cdot \text{School} + 50 \cdot \text{Aptitude}$$
$$+ 300 \cdot \text{Experience} - 2000 \cdot \text{Gendum}$$
$$- 200 \cdot \text{Gendum} \cdot \text{School} \qquad (2)$$

Using the same hypothetical data for the explanatory variables as before, I produced new values of earnings using equation (2) and (just for variety's sake) noise terms drawn from a distribution with standard deviation of 2,000. I then estimated a regression from the new data set, including the variable Gendum•School as an additional explanatory variable. The results are in table 4, where the variable "Interact" is simply Gendum•School.

Table 4. Multicollinearity illustration

Variable	"True value"	Estimated value	Standard error	t-statistic	Prob (2-tail)
Constant	5000.0	3500.2	2130.6	1.643	.108
School	1000.0	962.3	144.0	6.679	.000
Aptitude	50.0	61.2	12.3	4.966	.000
Experience	300.0	288.3	36.7	7.861	.000
Gendum	−2000.0	−4243.7	2916.5	−1.455	.153
Interact	−200.0	14.8	198.2	.075	.941

$R^2 = .909$

Observe that, in contrast to table 1, the coefficient for the gender dummy is now higher than the real value by more than a factor of two. The coefficient for the interaction term, by contrast, has the wrong sign and is close to zero. The other parameter estimates are not too far off the mark.

The poor results for the coefficients of Gendum and Interact are almost certainly a result of a severe multicollinearity problem. Note

that whenever Gendum = 0, Interact = 0 as well. Gendum is positive only when Interact is positive. We would expect a high correlation between them, and indeed the correlation coefficient is .96. Under these circumstances, it is no surprise that the regression cannot separate the effects of the two variables with any accuracy. The estimated coefficient for Interact is insignificant by any plausible test, and the coefficient for Gendum also has a large standard error that produces a rather poor t-statistic despite the high absolute value of the coefficient estimate.

Notwithstanding the considerable uncertainty introduced into the coefficient estimates, however, it is plausible that the multicollinearity problem here is not disastrous for an investigator interested in identifying the extent of gender discrimination. The reason is that the estimate of the *joint* effects of the Gendum and Interact may not be too far afield—one is inflated and one is understated, with the errors to a great extent canceling each other—and as a legal matter an estimate of the joint effect may be all that is needed. The caveat is that multicollinearity reduces the t-statistics for both variables, and might thereby lead the investigator to reject the hypothesis that discrimination is present at all. To deal with the effects of multicollinearity here, therefore, the investigator might simply wish to discount the low t-statistics, or else to omit one of the two variables and recognize that the coefficient estimate for the included variable will be biased and will include the effect of the omitted variable.[33]

In many instances, however, the investigator will not be satisfied with an estimate of the joint effect of two variables, but needs to separate them. Here, multicollinearity can become highly problematic. There is no simple, acceptable solution for all cases, though various options warrant consideration, all beyond the scope of this lecture.[34]

6. A final Note on the Law: Regression Analysis and the Burden of Proof

A key issues that one must confront whenever a regression study is introduced into litigation is the question of how much weight to

---

[33]It is important to recollect that this approach raises the problem of omitted variables bias for the other variables as well.

[34]See P. Kennedy, supra note 17, at 146–56.

give it. I hope that the illustrations in this lecture afford some basis for optimism that such studies can be helpful, while also suggesting considerable basis for caution in their use.

I return now to an issue deferred earlier in the discussion of hypothesis testing—the relationship between the statistical significance test and the burden of proof. Suppose, for example, that to establish *liability* for wage discrimination on the basis of gender under Title VII, a plaintiff need simply show by a preponderance of the evidence that women employed by the defendant suffer some measure of discrimination.[35] With reference to our first illustration, we might say that the required showing on liability is that, by a preponderance of the evidence, the coefficient of the gender dummy is negative.

Unfortunately, there is no simple relationship between this burden of proof and the statistical significance test. At one extreme, if we imagine that the parameter estimate in the regression study is the *only* information we have about the presence or absence of discrimination, one might argue that liability is established by a preponderance of the evidence if the estimated coefficient for the gender dummy is negative regardless of its statistical significance or standard error. The rationale would be that the negative estimate, however subject to uncertainty, is unbiased and is the best evidence we have.

But this is much too simplistic. Very rarely is the regression estimate the only information available, and when the standard errors are high the estimate may be among the least reliable information available. Further, regression analysis is subject to considerable manipulation. It is not obvious precisely which variables should be included in a model, or what proxies to use for included variables that cannot be measured precisely. There is considerable room for experimentation, and this experimentation can become "data mining," whereby an investigator tries numerous regression specifications until the desired result appears. An advocate quite naturally may have a tendency to present only those estimates that support the client's position. Hence, if the best result that an advocate can present contains high standard errors and low statistical significance, it is often plausible to suppose that numerous even less impressive

---

[35]See, e.g., Texas Department of Community Affairs v. Burdine, 450 U.S. 248 (1981).

results remain hidden, and conceivably shielded from discovery by the work product doctrine.[36]

For these reasons, those who use regression analysis in litigation tend to report results that satisfy the conventional significance tests—often the 5-percent significance level—and to suppose that less significant results are not terribly interesting.[37] Before most experts would feel comfortable asserting that gender discrimination has been established by a study such as that in our illustration, therefore, they likely would require that the coefficient estimate for the gender dummy be negative and statistically significant. Even then, they would anticipate a vigorous cross-examination based on a number of matters, many suggested by the discussion above.

Still more difficult issues arise when an exact parameter estimate is needed for some purpose, such as for computing damages. The fact that the parameter is "statistically significant" simply means that by conventional tests, one can reject the hypothesis that its true value is zero. But there are surely many other hypotheses about the parameter value that cannot be rejected, and indeed the likelihood that regression will produce a perfectly accurate estimate of any parameter is negligible. About the only guidance that can be given from a statistical standpoint is the obvious—parameter estimates with proportionally low standard errors are less likely to be wide of the mark than others.

Ultimately, therefore, statistics itself does not say how much weight a regression study ought be given, or whether it is reasonable to use a particular parameter estimate for some legal purpose or other. These assessments are inevitably entrusted to triers of fact, whose judgments on the matter if well informed are likely as good as those of anyone else.

---

[36]I will not digress on the rules of discovery here. In practice, the raw data may be discoverable, for example, while the expert's undisclosed analysis of the data may not be.

[37]See the discussion in Fisher, "Statisticians, Econometricians and Adversary Proceedings," 81 *J. Am. Stat. Assn.* 277 (1986).