Appendix: Basics of regression

Cliometrics is the explicit use of economic theory and measurement in the study of economic history. As the essays in this reader demonstrate, cliometrics has become a dominant method among economic historians. One of the central tools of cliometrics, and of all econometrics, is regression analysis. Multiple regression analysis is a means of fitting economic relationships to data. It lets us quantify economic relationships and test hypotheses about them.

Examine the data plotted in Figure A.1 for the Michigan furniture industry in 1889. Each dot on the scatter plot represents one worker. Clearly, there is a relationship between age and earnings. If we could draw a line through the scatter of points summarizing what is going on, the line would have an upward slope, since earnings generally rise with age. This is exactly what regression analysis does – it selects the line that provides a best fit to the data.

Estimating this line must begin with some economic theory. Most models of the labor market assume that wages are based on the value of the worker's contribution to output. Furthermore, in many jobs, productivity of young adults rises with maturity and experience. The theory, therefore, says that earnings depend on (rise with) age. Earnings are the dependent variable; age is the independent variable. It would be foolish to argue that the causation runs the other way, that age depends on or is caused by level of earnings. We can write this simple model mathematically as

$$\text{Earnings} = b_0 + b_1 \cdot \text{Age} + e$$

or more generally as

$$Y = b_0 + b_1 \cdot X + e$$

In the equation, b_0 and b_1 are the parameters or coefficients to be determined from the data. Here b_0 is the intercept term, and b_1 is the slope term in the equation of the line. The error term, e, represents the collective influences of any omitted variables that may also affect

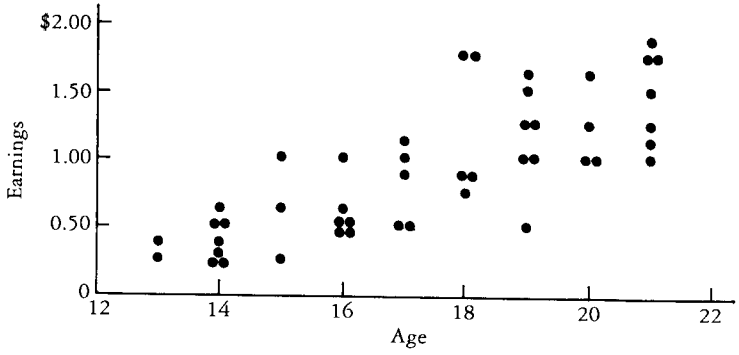

Figure A.1. Age and wage (dollars per day) Michigan furniture workers, 1889.

earnings. It recognizes that not every case will fall on the line.

Some criterion for a best fit is needed to choose values for the regression parameters, b_0 and b_1. The criterion most often used is to minimize the sum of squared differences between the actual values of Y and the fitted values for Y obtained after the equation line has been estimated. This is called the "least-squares criterion." If we denote the estimated parameters for the model by $\hat{b}_0$ and $\hat{b}_1$, then the fitted values for Y are given by

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 \cdot X$$

For each data point, the regression residual is the difference between the actual and fitted values of the dependent variable. The parameter values are chosen so that when all the residuals are squared and then summed, the resulting sum is minimized.

In the data from the scatter plot given previously, the result of fitting this equation using the least-squares criterion is

$$\text{Earnings} = -1.73 + 0.145 \cdot \text{Age}$$

where earnings are the dollars per day of furniture workers in Michigan in 1889 and age is in years. These data are used in Whaples (1992).

The slope parameter, $\hat{b}_1$, is of the most interest. It indicates that among young male workers below age twenty-two, a one-unit (i.e., one-year) increase in the age of the worker generally led to a $0.145 per day increase in earnings. The average daily earnings of this group was $0.89 (in 1889 dollars). Often the intercept is economically meaningless. This intercept literally tells us that a zero-year-old worker is predicted to have negative earnings. However, the intercept is important for fitting the best line through the data.

The beauty of computers is that they can easily think in many dimensions, so a more complicated model of earnings can easily be constructed, one that allows for many independent variables to influence the dependent variable and examines the impact of a change in each of the independent variables while holding the other independent variables constant. For example, the same data yield this regression line:

$$\text{Earnings} = -1.666 + 0.151 \cdot \text{Age} - 0.116 \cdot \text{Immigrant}$$

It indicates that, after controlling for age, immigrant workers earned \$0.116 less per day than nonimmigrant workers. Independent variables like this immigrant variable, which take on only two values (which = 1 if a condition holds and = 0 if it doesn't hold) are known as "dummy variables."

Our estimates of the true (but unknown) underlying parameters depend on the set of observations we use. We generally cannot observe every occurrence (e.g., only some farmers' account books have survived, or the 1910 U.S. census of population is so huge that only a small but representative fraction has been computerized). (The number of observations used in the regression's sample is often noted by the letter N.) If we could collect more and more samples and generate further estimates, the estimates of each parameter would follow a probability distribution. The probability distribution can be summarized by a mean and a measure of dispersion around that mean, a standard deviation referred to as the "standard error of the coefficient."

For any particular sample, least-squares estimates provide the best guess of the true underlying parameter. However, once we have information about the probability distribution for each coefficient, we can make statistical statements about our knowledge of the true parameters. Error terms are often assumed to be normally distributed. The normal distribution has the property that the area within 1.96 standard errors of its mean is equal to 95 percent of the total area. Thus, given our parameter estimate, $\hat{b}$, we can construct an interval around $\hat{b}$ within which there is a 95 percent probability that the actual parameter lies. That confidence interval is given by

$$\hat{b} + 1.96 \cdot (\text{standard error of } \hat{b})$$

Therefore, we should not only look at the point estimates of the coefficients, but also examine their standard errors. If the 95 percent confidence interval contains 0, then we cannot be certain that the true parameter b is different from 0. If we cannot be sure that the parameter is different from zero, then we cannot be sure that a relationship between the dependent and independent variables actually exists.

Regression printouts and tables report either the standard error of $\hat{b}$ or the t-statistic. The t-statistic is defined as

$$t = \frac{\hat{b}}{\text{standard error of } \hat{b}}$$

If the t-statistic is less than 1.96 in magnitude, the 95 percent confidence interval around $\hat{b}$ includes 0, and there is at least a 5 percent chance that b equals 0. We therefore say that our estimate is not statistically significant. On the other hand, if the absolute value of the t-statistic is greater than 1.96, we reject the hypothesis that $b = 0$ and call our estimate statistically significant.

Using the data on Michigan furniture workers again:

$$\begin{array}{rcccc} \text{Earnings} = & -1.666 & + 0.151 \cdot \text{Age} & - 0.116 \cdot \text{Immigrant} \\ & (0.306) & (0.017) & (0.087) \\ t = & -5.44 & 8.77 & -1.33 \end{array}$$

In this case, we are fairly confident (more than 95 percent confident) that earnings do climb with age, but we are not confident that immigrants actually earn any less (or more) than nonimmigrants. It is important to consider the "economic significance" of a coefficient in addition to its statistical significance. Suppose that the immigrant coefficient was statistically significant. Would the coefficient be economically significant as well? Would this 11.6 cent per day difference in earnings be economically important? This question cannot be answered by a computer. This is the historian's job.

The last thing to pay attention to in a regression table is the measure of the goodness of fit. This measures how much of the variation in the dependent variable has been explained by the variation in the independent variable(s). Goodness of fit is usually measured using the r-squared (R^2) or adjusted r-squared. This measure ranges from 0 (when the regression explains none of the variance) to 1 (when it explains all of the variance). A high R^2 by itself does not mean that the variables actually included in the model are the appropriate ones, because R^2 varies with the types of data being studied. Time-series data will often yield much higher R^2 values than cross-sectional data. In addition, if the theory underlying the equation is not valid, the equation could still yield significant coefficients with a high R^2, even though the results are meaningless. The adjusted-R^2 values from the preceding equations are 0.623 and 0.630. Adding the immigrant variable helps explain a small fraction of the variance in wages.

To summarize, these questions must be answered when examining a regression table:

1. What is the underlying theory? What is the dependent variable, and how is it assumed to be related to the independent variables?
2. What are the slope coefficients? How much does the dependent variable rise or fall when the independent variables change by one unit?
3. Are the slope coefficients statistically significant? Can we be sure that the coefficient is different from zero? Is the magnitude of the *t*-statistic greater than 1.96?
4. What is the R^2? How much of the variation in the dependent variable has been explained?

References

Robert Whaples, "Using Historical State Bureau of Labor Statistics Reports in Teaching," *Historical Methods*, 25 (Summer 1992), 132–6.