Statistical Evidence and Inference

Basic Methods of Analysis

Understanding the methods used by economists requires some basic terminology regarding the distribution of random variables. The mean of a distribution is simply the arithmetic average of all of the observations. The median is the observation that falls in the middle; half of the observations have values below the median and half are above. Both the mean and the median are measures of the center of the distribution, but the median is less sensitive to extreme observations. The variance is the average squared deviation from the mean for all of the observations, and is a measure of the spread of the distribution around the mean. The standard deviation is the square root of the variance.

The standard tool of economic data analysis is regression, which is a form of curve fitting. The most common method of regression used is Ordinary Least Squares (OLS). OLS can be used to fit any linear model of the form $y = a+bx+e$, where a is a constant and b is the coefficient that describes how y changes with changes in x. Because the relationship between x and y is not expected to hold exactly for every individual, there is an error term, e. For example, the relationship between age and hours worked (Figure 24) can be described as $hours = a + b * age + e$, and OLS used to obtain estimates of the parameters a and b.

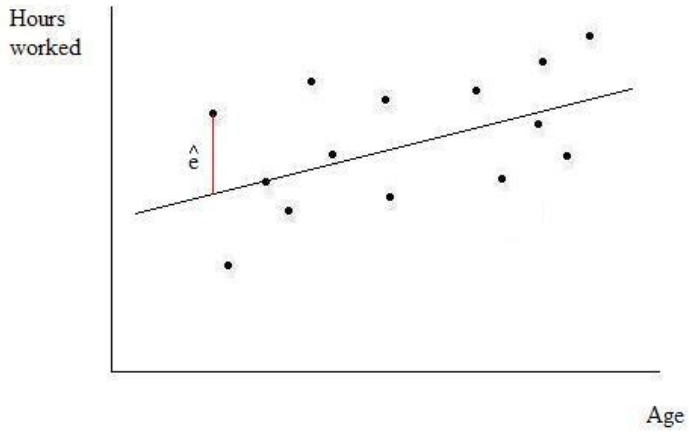

Figure 24

The ordinary least squares estimates are based on the assumption that the relationship between the two variables is linear; in this example, that the effect of age does not vary with age itself. The truth of this assumption cannot be verified apriori. The parameters are estimated "correctly" conditional on this linear model being the right one. The linearity assumption can be relaxed by

adding transformations of age, such as age squared, to the model. The addition of more explanatory variables transforms the model from a simple regression to a multiple regression. We could add health to the hours worked model, $hours = a + b * age + c * health + e$, to create a multiple regression model in which both health and age affect the number of hours worked by individuals. In this case, the coefficients on the explanatory variables describe the partial effect of the respective variable. The coefficient on age describes the effect of age on hours worked, holding health constant. The coefficient on health describes the effect of health holding age constant.

OLS estimates of the model's parameters are calculated by minimizing the sum of squared vertical deviations of each point from the fitted line, $\widehat{e}$. This method is greatly influenced by outliers, because of the squaring of the individual $\widehat{e}$. For example, a paper was written several years ago that found a very large effect of equipment investment on social rates of return by comparing investment and subsequent rates of growth across many countries. A scatterplot of the data looked like that in Figure 25, with most of the data clustered at the lower left and one outlier, which happened to be Botswana. Without Botswana in the sample, most of the effect disappeared. The effect of the outlier in this example was particularly large because the sample was relatively small. There are several ways to deal with the issue of outliers. In some cases, it is appropriate to discard the observations. In tax return data, there are often people with negative income; these observations are difficult to explain and are usually thrown out. It is also possible to change the weighting on outlying observations, by using bounded influence estimators which limit the effect of very unusual cases. In the end, the treatment of outliers is more of an art than a science. Statistical methods do not give guidance on when to ignore outliers; it is up to the researcher to decide if they are providing useful information or not.

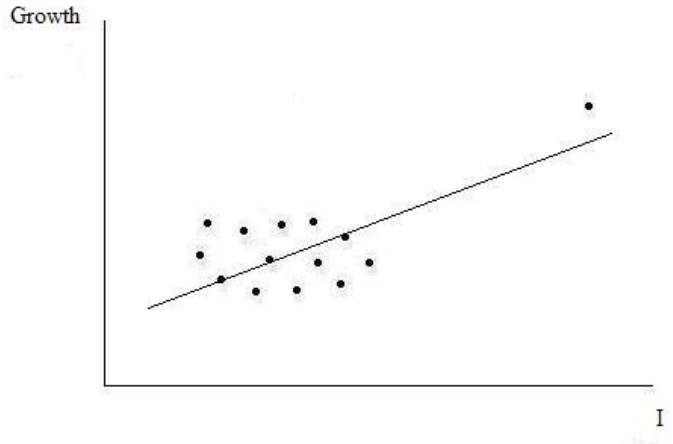

Figure 25

How can we know if the model estimated is the correct model? As mentioned above, the validity of a model's estimates is conditional on the model itself being the right one. The linearity of a model is not the only possible dimension on which the true model can differ from the estimated model, however. The set of explanatory variables must also be chosen. Hours generally rise with age, but so does the wage rate. If hours rise because the wage rate rises, but wages and age are also positively correlated, there will be a positive correlation between hours and age, even if age itself does not affect hours. If the model is estimated with age as the only explanatory variable, it will find that age affects hours worked. This is called omitted variable bias; because the true model includes the wage rate, the estimated model that leaves it out will not be able to generate good estimates of the included coefficients. If both wages and age are explanatory variables, the model will correctly attribute the effect of wages on hours worked and find that age has no effect. One way to test the choice of specification is to look at out-of-sample predictions. In a specific sample, it is always possible to fit a relationship to a set of points, or add more explanatory variables until model fits arbitrarily well. With out-of-sample prediction, we take estimates from one sample and see if the estimated model can predict the outcomes of another set of observations. For example, if we estimate the relationship between hours worked and age in a population with very similar wages, we may find no relationship between hours worked and age, simply because the wage does not vary much. If we try to use those estimates to predict hours worked in a population that has widely varying wages, the predictions will not be very good. The lack of stability of this model of hours worked and age across populations is an indication that the model is misspecified.

Another example from the tax literature may be more intuitive. Suppose that we want to explain the response of capital gains realizations to the tax rate, and we want to know if the tax rate itself affects realizations, or if only changes in the tax rate matter. This is the key empirical question in capital gains tax analysis. It is well-known that a cut in the capital gains tax rate will cause an increase in realizations in the short run. But we don't know if this is simply a timing effect, in which people shift asset realizations from one year to another, or if it is a permanent increase in realizations, which would represent more turnover of assets, less lock-in and distortion. This distinction is important for deadweight loss and revenue; if the second possibility is true, reducing the tax reduces deadweight loss. In the first case, there is no decrease in deadweight loss, only in revenue. If the revenue must be made up somewhere else, the tax cut could be welfare decreasing. The two possibilities are represented by two different models. If timing is important, the model of realizations is $r = \alpha + \beta t + \gamma \Delta t + \epsilon$; if only the level of the tax rate matters for realizations, the model is $r = \alpha + \beta t + \epsilon$. (Note that if we have only cross-sectional data on households and tax rates at a point in time, it is not possible to estimate the first model since no tax changes are available. Panel data, in which one household or individual is followed over a period of time, would be required.) If we estimate the model $r = \alpha + \beta t + \epsilon$, we will certainly find a large negative coefficient on the tax rate, but we cannot be sure that this truly represents the

effect unless we are sure that this second model is the correct one.

Another variation of omitted variable bias occurs when there are unobservable characteristics that affect the dependent variable. If we estimate a model of hours worked with the tax rate as the independent variable, but cannot control for the fact that some people simply prefer to work a lot, there will be a spurious correlation between hours worked and taxes. The people who have preferences for working a lot will have high hours, and the resulting high income will push them into higher tax brackets. The estimated coefficient on the tax rate will be positive, but the conclusion that high tax rates encourage work is false. In this case, we could use a change in the tax rate to isolate the response of hours to the tax rate from unobservable preferences.

In general, using the tax rate in a regression can be problematic because it is not exogeneous. An underlying assumption of the regression model is that causation runs only one way, from the independent to the dependent variable. However, in this example, the dependent variable, hours, also affects the independent variable, the tax rate. Instrumental variables regression can be used when this problem arises. To use IV, we need to identify some variable that does not depend on hours, but is still correlated with the marginal tax rate. The key to the technique is finding such a variable. In tax research, the tax rate on the first dollar of a specific type of income is often used as an instrument for the tax rate on the marginal dollar of that income. The tax on the first dollar is an attractive instrument because it is not affected by marginal decisions on capital gains realizations or hours worked, but it generally rises and falls with the tax on the marginal dollar.

Interpretation of Results

There are two ways in which the size or importance of an estimated coefficient can be evaluated. Statistical significance refers to the precision of the estimates. The empirical coefficients are estimates of the true parameters. Because the data are noisy, that is, because we do not expect the line to fit perfectly, we cannot recover α and β. Instead, we have estimates of them, called $\widehat{\alpha}$ and $\widehat{\beta}$. With the estimated coefficients come standard errors, which are a measure of uncertainty about the estimated coefficient, $\widehat{\beta}$. The standard error describes the spread of the distribution of the estimate $\widehat{\beta}$. The standard error is also a part of the calculation of the probability that the true coefficient is zero and the estimated coefficient is $\widehat{\beta}$. We say that the estimated coefficient is statistically significant if the probability that the true coefficient is zero conditional on the estimated coefficient is below some cutoff level, usually .05 or 5%. The closer the fitted line comes to the points, the lower the standard error and the more significant the coefficient. In Figure 26a, the data are relatively noisy, so it is hard to distinguish which of two fairly different possible fitted lines is the best. When the points are clustered closely together as in Figure 26b, we can be more confident in the estimated coefficients. Thus, noisy data can increase the standard errors of the model's estimates. If the independent variables are highly correlated with each other, leaving little independent variation to explain

changes in the dependent variable, this can also increase the standard errors. This multicollinearity is a common problem in empirical tax research. If a model of charitable contributions is estimated with the tax rate and taxable income as independent variables, the correlation between income and the tax rate in a progressive system will make estimation difficult. Sample size will also affect the standard error. With many independent observations, the standard errors shrink, so that even a small coefficient can be estimated relatively precisely.

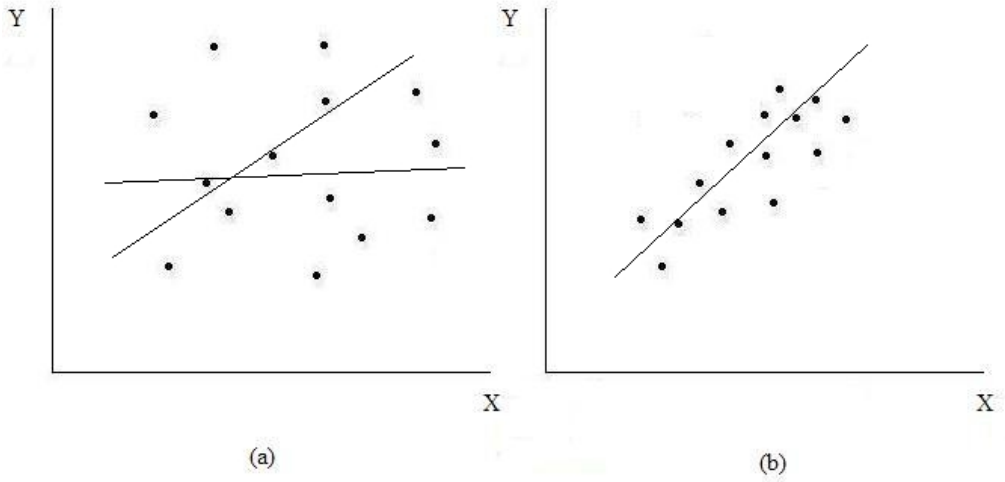

Figure 26

A statistically significant coefficient is not necessarily an economically important one. With a large enough sample size, we can be relatively sure that a coefficient is not zero, but it may still be very small. If we estimate the effect of a reduction in the capital gains tax from 30% to 20% to be an increase in realizations of 0.25% and find that the estimate is very significantly different from zero, the estimated effect is still very small.

Empirical research often reports other measures of the model's fit in addition to the coefficients. The R^2 is the percentage of the variation in dependent variable that is explained by the variation in the independent variables, or the percentage of variation in the dependent variable that is explained by the model. Because R^2, by definition, increases when additional variables are added, another measure, the adjusted R^2, is also frequently reported. This measure uses a "degress of freedom" correction, which penalizes the addition of more explanatory variables. Thus, the adjusted R^2 can be compared across models with different numbers of explanatory variables.

Every statistical conclusion in empirical research is based on the assumption that the model being estimated is the correct one. Thus, any interpretation of empirical results should begin with assessing the validity of the underlying model and its assumptions. In addition, it is important to note that statistical theory tells us that even if there is no relationship between two variables, one in twenty regressions will show a coefficient that is significant at the 5% level. Because the papers chosen for publication are nearly always those that come up with significant coefficients, the selection of research in journals is skewed and may present a false view of empirical relationships.