Econometrics for Economics Courses

Jeffrey Parker
Reed College

January 2017

Why economists need econometrics

Economic theories such as the supply-demand theory of competitive markets and the theory of rational consumer behavior can often tell us the general direction in which we should expect changes in economic conditions to influence decisions and economic outcomes. For example, the law of demand tells us that an increase in the price of cheese should (other things equal) decrease consumer purchases. However, such theories rarely answer the often equally important questions about the magnitude of these effects, for example, how much the quantity demanded will decrease if cheese becomes $0.20 per pound more expensive.

The sub-discipline of **econometrics** provides the link between economic theories and the real world. Econometric methods allow us to test whether observed data conform to the predictions of a theory and to estimate important parameters that measure the effects of variables on one another.

Narrowly speaking, econometrics is the application of statistical procedures to economic data. However, econometricians have developed many statistical models specifically for economic applications. Moreover, the practice of econometrics also requires an understanding of how economic data are measured, which functional forms tend to be most useful specifications for economic models, and other details involved in relating economic theory to observed data.

Why you need a little econometrics

Reed has two econometrics courses: Economics 311 for those who want to learn how to read and understand econometric applications in published literature and Economics 312 for those who want to learn to use econometric techniques. However, applied courses in every field of economics rely extensively on econometric tests and estimates. This brief and simple introduction is designed to make you familiar with a few of the most important concepts and methods of econometrics that you are likely to encounter frequently in your study of economics.

It introduces you to the concept of **linear regression**, which is the building block on which most of econometrics is based. Most econometric techniques are a variation of some kind on the use of linear regression to estimate and test hypotheses about a bivariate or multivariate relationship.

Simple regression: Fitting a line to a scatter of points

The simplest example of linear regression is the case of two variables where causality is known to run only in one direction. Suppose that economic theory tells us that local income in Portland *x* should have an effect on the quantity of parsnips purchased in Portland *y*. We assume that changes in *y* occurring for other reasons do not affect *x*, presumably because Portland citizens do not buy or sell enough parsnips to have a significant effect on their income.[1] In order to determine the magnitude of the effect of *x* on *y*, we will need to collect some observations on the two variables. Once we have a **sample** of **observation**—say, several years of corresponding values for *x* and *y*—regression analysis can be used to calculate and test hypotheses about estimates of the sensitivity of *y* with respect to changes in *x*.

The basic idea of linear regression is to fit a straight line to the collection of data points that we observe for *y* and *x*. A linear (straight-line) relationship between the variables can be represented by the equation

$$y_t = \beta_0 + \beta_1 x_t, \tag{1}$$

where β_0 and β_1 are unknown **parameters** whose values we wish to estimate.[2] These parameters define the nature of the linear relationship between *y* and *x*—whether it slopes upward or downward and how high or low the line lies. The parameter β_1 measures the effect on *y* of a one-unit change in *x*. This is the slope $\Delta y/\Delta x$ of the line representing the relationship. In terms of economic interpretation, we expect β_1 to be positive if parsnips are a normal good. The value of *y* at which the function intersects the vertical axis is given by β_0. A larger value of β_0 is associated with a line that lies vertically higher.

Regression with two observations

Suppose first that we have exactly two observations on *y* and *x*. In other words, we observe two independently generated pairs of values for the variables from different years. Let us call these two observations (x_1, y_1) and (x_2, y_2), where observation 1 is a measure of the two variables in year 1 and observation 2 is a measure of the variables in year 2.

If we plot these two "data points" on a graph with *x* on the horizontal axis and *y* on the vertical axis, we might get a diagram similar to the one in Figure 1, where the

[1] This assumption of "exogeneity" is extremely important. There are more advanced econometric technique that, in some cases, can be used if both variables affect one another.

[2] Of course, not all economic relationship are well approximated by straight lines. Linear regression analysis can be used with equal ease on any mathematical relationship in which the parameters enter in a linear way. For example, the logarithmic relationship $\log y = \beta_0 + \beta_1 \log x$ is one of the most commonly used functional forms in economics.

data points are labeled a_1 and a_2. As you can see, there is exactly one line that passes through the two data points. We shall represent the mathematical equation for this line as $y = b_0 + b_1 x$. (We are using b to represent out estimates of the true β parameters. Many authors is $\hat{\beta}$ in place of b.) The line is a "perfect fit" for the data, in the sense that both data points lie exactly on the line. In mathematical terms, $y_1 = b_0 + b_1 x_1$ and $y_2 = b_0 + b_1 x_2$. In the case of only two data points, fitting the best straight line to the data is easy! The slope of this line is b_1, which is our empirical estimate of β_1, while the value of y where the best-fit line intercepts the y axis is b_0, our estimate of β_0.

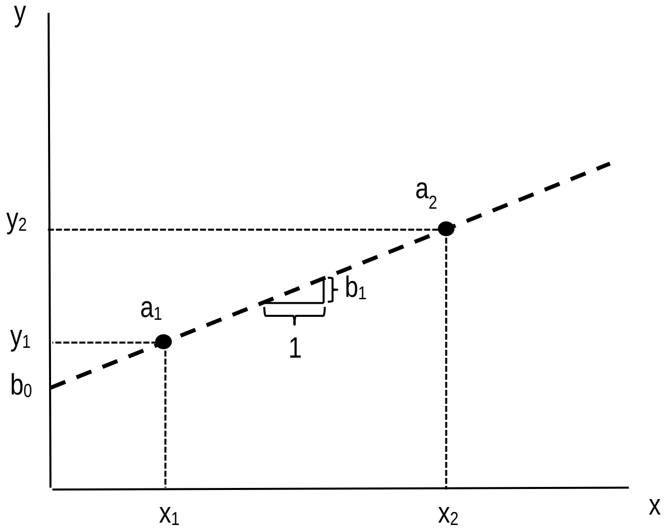

Figure 1. Best-fit line with two data points

Adding a third observation

Suppose now that we obtain a third data point (x_3, y_3) by observing a third year. Should we expect that this data point would lie exactly on the line connecting the first two points? If the demand curve of equation (1) holds precisely for all three observations, then all three should obey the same linear relationship and they *should* be on the same line.

However, measured economic relationships are never that precise. For one reason, variables are observed with error. For another, the relationship between any two variables is usually subject to disturbances by additional variables that are not included in the equation (and often by variables whose values cannot be observed at all). Consequently, econometricians usually interpret the hypothesis of a linear relationship to assert that all of the data points should lie *close to* a straight line.

However, it would be very unusual for the added data point to lie *exactly on* the line that passed through the first two.

In order to allow for this “imperfection” in our two-variable linear relationship, we add a disturbance term or error term to equation (1). The resulting equation looks like

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \tag{2}$$

where ε_t is the **disturbance term**, which is usually modeled as a **random variable**.[3]

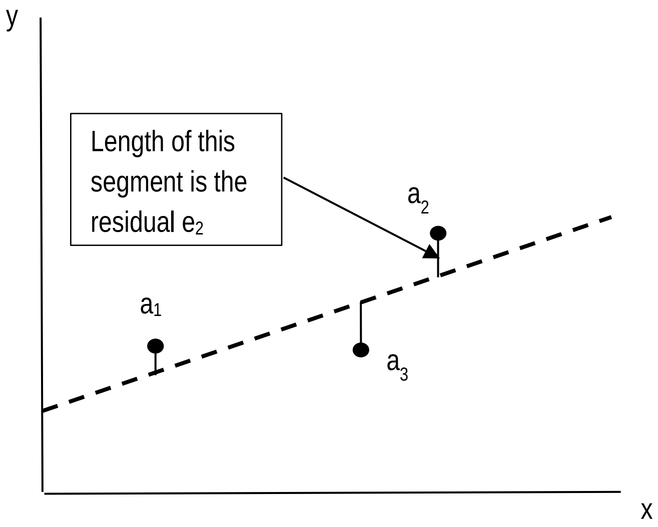

Figure 2. Best-fit line with three data points

Suppose that the three data points are as shown in Figure 2, so that they do not line up on the same straight line. Now there is no single line that fits all three data points exactly. What criterion should we use to select which line best fits the three data points? In order to answer that question, we must first choose a method to measure “how close” any particular line lies to the collection of three points, and second find and choose the line that lies "closest" to the points according to that measure. The measure most often chosen is that of **least-squares**, and the line that is chosen as the best-fit line is the one that minimizes the squares of the vertical

[3] A variable is considered random if we assume nothing about how it is determined except that it follows a given probability distribution, meaning that it takes on particular values with a probability that is known or can be estimated. The most common random variables in econometrics follow the normal probability distribution, which means that the likelihood that they take on particular values is given by a “bell curve.”

distances of the three points from the line. In Figure 2, the short vertical line segments signify the **residuals**—the vertical deviations of the observed points from the best-fit line. If we again denote the best-fit values of β_0 and β_1 by b_0 and b_1, then the residual for observation t is $e_t = y_t - b_0 - b_1x_t$.[4]

Some of the residuals are positive—those for observations where the actual value of y_t lies above the best-fit line such as observations 1 and 2 in Figure 2—and some are negative (observation 3 in Figure 2, where the point lies below the line). Therefore, we cannot simply minimize the sum of the residuals. If we worked with the sum of the residuals, the positive and negative residuals would cancel out. In order to avoid this canceling, we *square* each of the residuals (since the square is positive whether the residual is positive or negative) and choose as our best-fit line the one that minimizes the sum of the *squares* of the residuals. The best-fit line we determine by this criterion is called the **least- squares regression line**. Introductory econometrics texts give formulas for calculating the values of b_0 and b_1 for the best-fit line, but you need not be concerned with the precise method of calculation. Many computer programs exist to perform these calculations for you.

Observations and degrees of freedom

Before we leave the two-data-point and three-data-point examples, there is one additional concept that can be introduced. With two observations, there is only one line that makes sense as a best-fit line. Statistically, we would say that there were no **degrees of freedom** in the choice of lines. Adding the third data point means that there would generally be at least one point that is off the regression line (unless the three happened to line up exactly). This one "extra" point beyond the two needed to define a line gives us one "degree of freedom" in choosing the line.

If there were four data points, then we would have two degrees of freedom—two additional points beyond the two that are required to define a line. In general, a two-variable regression has $N - 2$ degrees of freedom, where N is the number of observed data points in the sample. Degrees of freedom have important uses in the testing of hypotheses about the regression line.

Three variables and three dimensions

Very few economic relationships can be adequately characterized by just two variables. For example, the demand for parsnips in Portland may be thought of as a function of Portlanders' incomes, but it is also surely affected by many other variables such as the price of parsnips and the prices of eggplants and other substitutes. Economists are fortunate that the case of simple regression can be easily

[4] Be careful to notice the distinction between the disturbance term ε_t and the residual e_t. The disturbance term is the deviation of observation t from the line representing the *true* relationship between the variables: $\varepsilon_t = y_t - \beta_0 - \beta_1x_t$.

generalized to incorporate more than two variables. A regression equation with more than one explanatory variable on the right-hand side is a **multiple regression**.

Suppose that we generalize our parsnip demand curve so that the quantity of parsnips demanded is assumed to be influenced not only by income but by the price of parsnips *p*.[5] Including an error term, we could write this relationship as

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 p_t + \varepsilon_t. \tag{3}$$

How would we represent an equation such as (3) geometrically? We need a picture that shows how *y* changes when either *x* or *p* changes. In order to produce such a picture we must use three dimensions, with *y* measured vertically and *x* and *p* measured in two horizontal dimensions. The three-dimensional analogy to the two-dimensional regression line is a three-dimensional "regression plane." Just as with the two-dimensional case, there is a true regression plane that represents the (unknown) data-generating process and an estimated regression plane that is our best-fit estimate. The error term for observation *t* (ε_t) is the vertical distance between any point and the *true* regression plane, while the corresponding residual $e_t = y_t - b_0 - b_1 x_t - b_2 p_t$ is the vertical distance from the point to the *estimated* regression plane.

Following the logic of the least-squares estimator for the simple regression (two-variable) case, the least-squares estimator for the multiple regression model is the plane that minimizes the sum of the squared residuals for our sample of observations.

Mathematically, it is easy to extend least-squares estimation to accommodate more than two explanatory variables and three dimensions. However, pictorial representation of higher-order models is doomed by our inability to visualize more than three dimensions. Modern computer programs are able to find the coefficients for least-squares regression "hyperplanes" involving 100 or more dimensions (variables).

Measuring goodness of fit

One obvious question that we might ask about our estimated regression line is how closely it fits the data. The most common measure of goodness of fit is the R^2 statistic. This measures the share of the variation in the variable we are predicting that is explained by the estimated effects of the right-hand variable(s). In our parsnip regression with income and price on the right-hand side, if the R^2 value was 0.83, then we would conclude that 83% of the sample variation in quantity consumed was explained by the linear function of income and price on the right-hand side. The remaining 17% is unexplained and attributed to the disturbance term. A higher value

[5] Note that we must still assume that the price of parsnips in Portland is not affected by Portland demand, perhaps because Portland is a small part of a national market with a homogeneous price.

of R^2 indicates a better fit, since more of the variation is explained by movements along the estimated regression line and less by deviations from the line.

Accuracy of the least-squares estimator

Suppose our estimate b_2 is –80, meaning that the estimated effect of a 1 cent per pound increase in the price of parsnips is to reduce quantity demanded by 80 pounds per year. Is this effect "significant"?

That question can be answered in two different ways. First of all, –80 represents our best available estimate of how much a price change will affect quantity consumed. Whether this effect is a little or a lot depends on the economic context in which it occurs. Depending on costs and other conditions, it may be small enough effect that the big grocery-store chains in Portland decide to raise the price of parsnips to increase their profit margins. Alternatively, it may be large enough that grocers will resist a price increase as reducing sales volume by too much. To assess whether the coefficient's magnitude is "economically significant" in that sense requires more knowledge of the economic setting of the demand equation than we have given in this example.

However, when econometricians assess the "significance" of an estimated coefficient they are usually talking not about the estimated coefficient's economic implications but about its **statistical significance**. Our least-squares estimates are just that: estimates. Even good estimates are invariably imperfect, either a little too high or a little too low. How confident can we be that the –80 we estimated for β_2 should not be –60, –120, or even zero or +40?

Reliable estimates are ones that we think have a high probability of being close to the true parameter value—in this case, to the actual, real-life sensitivity of quantity consumed to price. This is usually assessed based on two criteria relating to how the estimator would perform in a hypothetical experiment where we had available many independent samples to estimate the parameter: (1) Across many samples, would the estimator we are using have the right value on average? (2) How widely would the values of the estimator (across the repeated samples) vary around the true parameter value?

Estimators that are correct on average across many samples, though not necessarily for any particular sample, are called **unbiased estimators**. The dispersion across samples of the estimates is measured by their **variance**. A larger variance means that the estimates are more widely dispersed (and less desirable).

Our assessment of the reliability of an estimator is made using its **standard error**, which is an estimate of the square root of its variance. A small standard error indicates a relatively precise estimate; a large standard error shows that our estimate is likely to be subject to large errors.

The standard error of an estimator is determined largely by two considerations. The first is the degree to which the available observations conform to the best-fit line. If the observations are all tightly clustered along the line, this gives us considerable confidence about the values of the slope and intercept parameters, leading to small standard errors. Second, for a given degree of fit, the more observations we have the smaller is the standard error. It gives us more confidence in our estimates if there are 200 observations tightly arranged near the line than if there only 20. Thus, a sample with many observations that fit the line well will lead to precise estimates of the parameters (small standard errors) while a small sample where the observations deviate substantially from the line gives less precise estimates (larger standard errors).

Using the least-squares estimator to test hypotheses

For many estimators, an interval that is centered on the true parameter value and that extends a distance of two times the standard error in each direction will include the calculated value of the estimator 95% of the time (in repeated samples). We use this property to create a 95% **confidence interval** for our parameter. Suppose that our calculated slope estimate of –80 has a standard error of 15. Two times the standard error is 30, so we can be 95% confident that the true parameter value for the slope lies within 30 units on either side of –80, *i.e.*, inside the range (–110, –50). If, perhaps by obtaining additional observations, we were able to reduce the standard error from 15 to 10 (without changing the "point estimate," which stays at –80), then our 95% confidence interval would narrow to (–100, –60).

Confidence intervals allow us to make probabilistic statements about our estimates. For example, with an estimate of –80 and a standard error of 15, we can be pretty confident that the true value of the demand slope parameter is not zero, since zero does not fall inside (or even close to) the confidence interval. We would feel less sure about rejecting –65 as a possible value for β_2, however, since it is contained within the confidence interval.

The exact same process we used to form confidence intervals is used in reverse to perform **tests of hypotheses**. We do this when we are interested in the answer to a question such as "Is it plausible that this sample of observations is generated by a process where the true slope of the demand function is *r*?" The affirmative answer to this question is called the **null hypothesis**. For example, we are often interested in the question of whether the estimated coefficients allow us to conclude that the demand curve slopes downward. We can cast this as a hypothesis test by formulating the null hypothesis that $\beta_2 = 0$ and determining whether the data stray sufficiently from what we would expect if that hypothesis were true that we can statistically **reject the null hypothesis** with, say, a 95% probability of being correct. It turns out that our statistical decision can be determined from the 95% confidence interval. It the value that the parameter is assumed to have under the null hypothesis (zero, here) lies within the confidence interval, then our sample is deemed to be consistent with the null hypothesis and we cannot conclude decisively based on our sample that $\beta_2 < 0$. In the case discussed above, however, zero was well outside of the confidence interval. Thus, in this case, we reject the null hypothesis that $\beta_2 = 0$ and conclude that the slope of the demand curve is negative.

Hypothesis tests are often presented in the form of a ***t*-statistic** or an associated **probability value** (*p*-value). The *t*-statistic takes its name from the probability distribution that applies to many test statistics in econometrics. We calculate the *t*-statistic for the test of the null hypothesis that the parameter is zero by dividing the estimate of the parameter by its standard error. For example, with a parameter estimate of –80 and a standard error of 15, the associated *t*-statistic would be –80/15 = –5.33. The decision rule for our hypothesis test then becomes rejecting the null hypothesis that the parameter is zero if the absolute value of the t-statistic is larger than some "critical value," which (for a 95% level of confidence) is usually close to two. Since 5.33 is much larger than 2, we reach the same conclusion here that we did above: it is unlikely that our sample came from a world in which price has no effect on quantity demanded.

Although 95% is a very common confidence level at which to perform hypothesis tests, one can choose a higher or lower significance level as well. For example, even if we cannot be 95% sure that a parameter is not zero (*i.e.,* we cannot reject the null hypothesis of a zero parameter at the 5% significance level) we may still feel quite sure of our result if we can be 94% or even 90% confident.[6] Many computer programs that are used for regression analysis make it very easy to know the exact level of confidence that can be associated with a statistical test. These programs report a *p*-value for each estimated coefficient. This value is the smallest significance level at which the null hypothesis can be rejected, which is one minus the largest confidence level at which we reject the null. Thus, a reported *p*-value of 0.035 for a test of whether a coefficient is zero would indicate that zero lies just on the border of a 1 – 0.035 = 0.965 (96.5%) confidence interval for the parameter. If we choose a confidence level higher than 96.5%, say, 99%, (a significance level smaller than 3.5%, say, 1%) then we cannot reject the null hypothesis. In this example, we cannot be 99% certain that the data did not come from a world with a true parameter value of zero, but we can be 95% certain of this.

A sample regression table

In reading regression results in a published paper, there will usually be a regression table that looks something like the one shown below.[7] The variables shown in the left-hand column as the right-hand regressors in the equation. The three columns refer to three different econometric estimation procedures that are described in the text of the paper.

[6] Note the distinction between two closely related numbers. The "confidence level" is the level of certainty at which we desired to know that the null hypothesis is false, while the "significance level" is the smallest probability we wish to allow of the null hypothesis being true. The confidence level is equal to one minus the significance level.

[7] From Diego Useche. 2014. "Are Patents Signals for the IPO Market? An EU–US Comparison for the Software Industry." *Research Policy* 43 (8):1299-1311.

Table 5
Patent applications and the amount of money collected at IPO.

Variables	1 EU–US OLS	2 EU–US HECKMAN 2S2	3 EU–US GMMEUUS
PATAPPUS	0.00338**	0.00341**	0.00507***
	(0.00150)	(0.00145)	(0.00137)
PATAPPEU	0.0134***	0.0108***	0.0113***
	(0.00349)	(0.00378)	(0.00378)
FCITATIONSUS	0.000906	0.00002	−0.000155
	(0.00132)	(0.00253)	(0.00150)
FCITATIONSEU	0.00975	0.0232	0.00996
	(0.0341)	(0.0146)	(0.0334)
LOG (TOTAL ASSETS)	0.622***	0.535***	0.614***
	(0.0401)	(0.0477)	(0.0386)
LOG (SALES TO ASSETS)	0.163***	0.155***	0.159***
	(0.0459)	(0.0522)	(0.0442)
VCUS	0.397***	0.633**	0.360***
	(0.127)	(0.309)	(0.123)
VCEU	0.483***	0.609*	0.508***
	(0.167)	(0.325)	(0.160)
CORPVCAP	0.0407	0.0404	0.0807
	(0.169)	(0.191)	(0.163)
LOG (AGE AT IPO)	0.00136	0.0639	−0.00457
	(0.0609)	(0.0991)	(0.0579)
NEW MARKET	0.0353	−0.261*	0.0378
	(0.116)	(0.150)	(0.111)
SOFT_RATIO	3.921**	1.370	4.296***
	(1.546)	(2.304)	(1.374)
LOG(PERCENT SOLD)	0.501***	0.283***	0.482***
	(0.0927)	(0.104)	(0.0885)
EU	−0.673***	−1.278***	−0.675***
	(0.142)	(0.335)	(0.136)
Financial ratios	Yes	Yes	Yes
Annual Dummies	Yes	Yes	Yes
Intra-industry dummies	Yes	Yes	Yes
Country dummies	Yes	Yes	Yes
Constant	1.976**	4.229***	1.948***
	(0.823)	(1.055)	(0.753)
Mills		0.984**	
		(0.491)	
Observations	476	476	476
Adjusted R^2	0.758		0.756

Robust standard errors in parentheses.

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

Focus on column (1). The number shown in each row is the estimated coefficient of that variable in the equation. For example, the variables log (total assets) has an estimated coefficient of 0.622, meaning that an increase of one unit in that variable causes an increase of 0.622 units in the dependent variable. Below the coefficient is its standard error, shown in parentheses. The asterisks to the right of some of the coefficients indicate the level of statistical significance that we can attach to the coefficient. As clarified by the note below the table, three asterisks means that we can reject the null hypothesis that the coefficient equals zero at the 1% level of significance.

At the bottom of the table, the number of observations and the (adjusted) R^2 are shown. The rows with "Yes" in the table refer to the presence of various sets of control variables whose coefficients are not of direct interest but which are included to control for other possible sources of variation in the dependent variable. This sort of entry is common when there are so many such variables that the full table with all of their coefficients wouldn't fit on one page.

Sometimes a cell in the table will be empty, although this is not the case in this particular table. That means that the variable in that row was omitted from the regression being shown in that column. Sometimes we are not sure which variables should be included and we present alternative models with and without certain variables so that the reader can compare the results.

Assumptions of least-squares regression

In the previous sections, we considered mathematical and geometric interpretations of the problem of finding the best fit for a set of data points. We asserted that the method of least-squares was the technique most often used to estimate such a best-fit line (or plane or hyperplane), and we examined some of the statistical properties of this estimator under a set of assumptions. In practice, the choice of estimation method is not so simple. In this section, we examine briefly the assumptions that one must make about the underlying data-generating process in order for the **ordinary least-squares** (OLS) estimator to be optimal. For some cases in which the OLS estimator is not desirable, econometricians have devised better estimators. (You can study all about these by taking Econ 312!)

In order for the OLS estimator to have desirable properties such as unbiasedness and efficiently low variance, the error terms of the data-generating process must obey several properties. First, the error term of each observation must follow an identical probability distribution (usually assumed to be a normal distribution). Second, the error terms must be statistically independent of one another, so that the value of the error term of one observation is not correlated with the values of the error terms of other observations. Third, the error term must be statistically independent of all of the variables on the right-hand side of the equation (the regressors). In other words, a positive error term must be equally likely when any given explanatory variable has a small value as when it is large.

Under these assumptions, the **Gauss-Markov Theorem** demonstrates that the OLS estimator has the smallest variance (or standard error) of all linear, unbiased estimators. We therefore say that under these conditions OLS is BLUE, which stands for **Best Linear Unbiased Estimator**. However, when one or more of these conditions is violated, OLS is no longer BLUE and we should consider other estimation techniques.

A common violation of the first assumption is the case of **heteroskedasticity**, which occurs when some observations have error terms that tend to be more variable than others. For example, in many macroeconomic applications researchers have found that observations after 1974 tend to be distributed more widely around the fitted regression line than observations from 1973 and before. One "eyeball" test for heteroskedasticity is to look at a plot of the residuals from the regression. If there are some intervals of time in which the residuals are far away from zero and others in which the residuals cluster very close to zero, then heteroskedasticity may be a problem. Of course, any random sample will have *some* variation from period to period in the magnitude of the residuals. Statistical tests can be used to determine whether the degree of variation in residuals is large enough that it is unlikely to have occurred through such random variation.

If heteroskedasticity is present, then the OLS estimator is still unbiased (correct on average), but it is no longer the best linear unbiased estimator. Moreover, the estimated standard errors from an OLS regression are biased, so that statistical tests based on these OLS standard errors are not valid.

Autocorrelation occurs when there is correlation across observations in the error terms. In a time-series context, this is often called **serial correlation**. Most time-series models have serious problems with serial correlation because if a substantial positive (negative) disturbance occurs in period *t*, it is usually the case that the disturbance in period $t + 1$ is likely to be positive (negative) as well. As with heteroskedasticity, there are both tests and corrections for problems of autocorrelation. An "eyeball" test for autocorrelation involves looking for patterns of several positive residuals in a row and several negative residuals in a row. If the residuals usually tend to have the same sign for several consecutive observations, then serial correlation is likely.

As with heteroskedasticity, the OLS estimator is still (usually) unbiased in the presence of serial correlation, but it is not the best estimator and the OLS standard errors are incorrect, which makes the usual confidence intervals and hypothesis tests invalid.

Econometricians encountering problems with heteroskedasticity or autocorrelation have two choices: (1) use a "generalized least-squares" model that corrects for the problem with an estimator that is optimal under those circumstances, or (2) use the sub-optimal OLS estimator but correct the standard errors for bias so that valid confidence intervals and hypothesis tests are possible. The latter procedure is far more common today. This is what is being used if you see an author refer to **robust standard errors**.

Endogeneity bias and instrumental variables

Violation of the third assumption described above—the lack of independence between the regressors (right-hand variables) and the disturbance term—is in many ways by far the most serious. This occurs whenever a shock to the dependent variable through the disturbance has an effect on one of the supposedly independent variables of the regression, so that the right-hand variable in question is **endogenous** rather than exogenous. In this case, OLS estimators are biased. Worse yet, even in large samples they will not tend to converge on the true values of the parameters.

This **endogeneity bias** is difficult to detect using statistical tests, so one often must use economic theory and reasonable judgment in choosing a model where the regressors are likely to be truly exogenous. If regressors are unavoidably endogenous, then one must typically use an **instrumental variables** (IV) estimator.

Instrumental variables models rely on finding one or more **instruments** effectively to replace each endogenous right-hand variable in the regression (each one that is correlated with the disturbance). Valid instrumental variables, or instruments, must satisfy three conditions: (1) they must be exogenous (uncorrelated with the disturbance term and unaffected by changes in the dependent variable), (2) they must be strongly correlated with the endogenous regressor that they are intended to replace, and (3) they must not have an effect, on their own, on the dependent variable.

If a valid instrument can be identified, then instrumental-variables regression replaces the endogenous variable on the right-hand side of the regression with the best prediction we can make for that variable based only on the instruments (and other exogenous variables). What we are doing is decomposing the endogenous regressor into two parts: an exogenous part that is based only on the exogenous regressors and the exogenous instrument(s) and an endogenous part—what is left in the variable after the exogenous prediction is removed. By using only the exogenous part in the IV regression, we avoid the problem of endogeneity bias and have valid coefficient estimates.

IV regression is very common, but often very controversial. The results are valid only if all three of the instrument assumptions above are valid. One can often question the exogeneity of the instruments, the strength of the correlation with the endogenous regressor, or the absence of an independent effect of the instrument on the dependent variable. This leads to frequent debates about how reliable any given IV regression result is.

A quick sketch of some common time-series models and concepts

There are three main kinds of data samples in econometrics: cross-section (where we observe a collection of individual units at the same time), time-series (where we have observations on a single unit at many dates), and pooled data which combine both of these. Most of what has been discussed above applies in all three settings. In this section and the next one we consider some common models that apply to time-series data and to special kinds of pooled-data samples.

One problem in time-series data is that most economic effects are not immediate but instead are spread over time. If x has an effect on y, it is likely that a change in x_t will not only affect y_t but will also have some **lagged effect** on y_{t+1}, y_{t+2}, and so on. Thought of another way, the current value of the dependent variable y_t depends not just on the current x_t but also on lagged values x_{t-1}, x_{t-2}, … We can often estimate such dynamic effects with **distributed-lag models** in which we include lagged values of x on the right-hand side as regressors.

Another common problem in time-series samples is **nonstationarity**. Basic OLS regression is only valid when the variables are "stationary." This means, roughly, that they have the same probability distribution for every date in the sample. This is obviously problematic for variables such as GDP, which grow systematically over time. The distribution of GDP for 2017 is not the same as for 1997—it is centered around a much higher mean (average).

One way of dealing with nonstationary is to **difference** the variables, running the regression in terms of changes (or growth rates) of the variables rather than levels. Differencing often makes a nonstationary variable stationary, so the differenced regression may be valid even if the original regression is not.

A special case arises with two or more variables are **cointegrated**. This means that they are both nonstationary (for example, both growing over time), but they are nonstationary "together," so that there is a stable (stationary) relationship between them over time. For example, house prices in Portland and house prices in Beaverton are both nonstationary because they grow over time. But they may be cointegrated if there is a stable, long-run relationship between them. Econometricians have developed estimators that can be used in cointegrated models.

A final estimation method that is very common in macroeconomics is the **vector autoregression** (VAR). The VAR is a method of estimating the relationships among a set of variables without making strong (and perhaps invalid) exogeneity assumptions. A VAR model has an equation for each of the set of variables in which that variable is regressed on lagged values (up to some specified maximum lag) of all of the variables in the set. Because lagged values can usually be assumed to be exogenous, the VAR avoids the problem of endogeneity and can be estimated by OLS.

There are several uses of VARs. They can easily be used for forecasting the future behavior of the set of variables. They can also be the basis for **Granger causality tests**, which attempt to assess the causal relationships among the variables under the assumption that x causes y if and only if knowing previous values of x helps you predict y, taking into account y's own previous history.

VARs can also be used to estimate **impulse-response functions** (IRFs), which are estimates of the effect that a current shock to one of the variables would have on all of the variables of the system. IRFs are extremely useful, but require additional "identifying assumption" that are analogous to assuming exogeneity. Since these identifying assumptions may be incorrect, it is possible to get invalid IRFs even if the VAR is totally satisfactory.

Fixed-effects estimation of panel-data models

One of the best ways of estimating economic relationships is by using a data set that varies both over time and across units. For example, using data for a sample of years for all 50 U.S. states takes advantage both of variation across states and variation over time to estimate the relationship. When we have a pooled cross-section time-series sample in which we have observations on the same units over time (such as states), we call it **panel data**.[8]

One problem with using panel data is that there will always be regression variables that we cannot measure. These variables are likely to be correlated state-to-state with some of the variables that we have included in the regression, which leads to **omitted-variables bias**. One way to try to eliminate this bias is to include **dummy variables** for each state in the regression.[9] The dummy variables "soak up" the variation that is strictly across states and allow the coefficients to measure variation across time in all of the states. Specifically, the coefficients on the regression variables do not measure why Oregon's values of the dependent variable are different from California's they only measure why the change in Oregon's values from year to year differs from the change in California's.

When dummy variables are included for each cross-sectional unit, we call it a **fixed-effects regression**. The coefficients on the dummy variables measure the "fixed" effect of being in a particular state. We sometimes also include **time fixed effects**, or dummy variables for each time period, to capture variation that is common to each state over time. Regressions with both unit and time fixed effects require a high standard of correlation between the variables in order to find significance. If the relationship between y and x is purely between the states or purely over time, the fixed-effects model will attribute it to the dummy variables. The only part of the relationship between y and x that will be captured in the regression coefficient on x is differences across states in the differences (changes) over time. This is an example of a **differences-in-differences** estimator.

What you know and what you do not know

This brief summary has provided you with a simple description of linear regression and a few common variations. These are the most basic methods of

[8] In contrast, consider a survey taken every years, but where the individuals surveyed are not the same in each year. That would be time-series cross-section data, but not panel data.

[9] Dummy variables are variables that take on only the values zero and one. For example, a dummy variable for Oregon would have the value 1 for Oregon's observations and 0 for all other states' observations.

econometric analysis. Much has, of course, been skipped. If you are interested in learning more about these methods, any basic econometrics text can provide you with the details. For the moment, this introduction should allow you to read some basic econometric studies and have a pretty good idea what is going on.