APPENDIX

The Basics of Regression

This appendix explains the basics of **multiple regression analysis**, using an example to illustrate its application in economics.[1] Multiple regression is a statistical procedure for quantifying economic relationships and testing hypotheses about them.

- **multiple regression analysis** Statistical procedure for quantifying economic relationships and testing hypotheses about them.

In a **linear regression**, the relationships are of the following form:

$$Y = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e \qquad \textbf{(A.1)}$$

- **linear regression** Model specifying a linear relationship between a dependent variable and several independent (or explanatory) variables and an error term.

Equation (A.1) relates a *dependent* variable Y to several *independent* (or *explanatory*) variables, $X_1, X_2, \ldots$. For example, in an equation with two independent variables, Y might be the demand for a good, X_1 its price, and X_2 income. The equation also includes an *error term* e that represents the collective influence of any omitted variables that may also affect Y (for example, prices of other goods, the weather, unexplainable shifts in consumers' tastes, etc.). Data are available for Y and the Xs, but the error term is assumed to be unobservable.

Note that equation (A.1) must be linear in the *parameters*, but it need not be linear in the variables. For example, if equation (A.1) represented a demand function, Y might be the logarithm of quantity ($\log Q$), X_1 the logarithm of price ($\log P$), and X_2 the logarithm of income ($\log I$):

$$\log Q = b_0 + b_1 \log P + b_2 \log I + e \qquad \textbf{(A.2)}$$

Our objective is to obtain *estimates* of the parameters $b_0, b_1, \ldots, b_k$ that provide a "best fit" to the data. We explain how this is done below.

AN EXAMPLE

Suppose we wish to explain and then forecast quarterly automobile sales in the United States. Let's start with a simplified case in which sales S (in billions of dollars) is the dependent variable that will be explained. The only explanatory variable is the price of new automobiles P (measured by a new car price index scaled so that 1967 = 100). We could write this simple model as

$$S = b_0 + b_1P + e \qquad \textbf{(A.3)}$$

[1]For a textbook treatment of applied econometrics, it's hard to think of a better reference than R. S. Pindyck and D. L. Rubinfeld, *Econometric Models and Economic Forecasts*, 4th ed. (New York: McGraw-Hill, 1998).

In equation (A.3), b_0 and b_1 are the parameters to be determined from the data, and e is the random error term. The parameter b_0 is the intercept, while b_1 is the slope: It measures the effect of a change in the new car price index on automobile sales.

If there is no error term, the relationship between S and P would be a straight line that describes the systematic relationship between the two variables. However, because not all the actual observations fall on the line, the error term e is required to account for omitted factors.

ESTIMATION

In order to choose values for the regression parameters, we need a criterion for a "best fit." The criterion most often used is to *minimize the sum of squared residuals* between the actual values of Y and the *fitted* values for Y obtained after equation (A.1) has been estimated. This is called the **least-squares criterion**. If we denote the estimated parameters (or *coefficients*) for the model in (A.1) by $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$, then the *fitted* values for Y are given by

• **least-squares criterion** Criterion of "best fit" used to choose values for regression parameters, usually by minimizing the sum of squared residuals between the actual values of the dependent variable and the fitted values.

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \cdots + \hat{b}_k X_k \qquad \textbf{(A.4)}$$

Figure A.1 illustrates this for our example, in which there is a single independent variable. The data are shown as a scatter plot of points with sales on the vertical axis and price on the horizontal. The fitted regression line is drawn through the data points. The fitted value for sales associated with any particular value for the price values P_i is given by $\hat{S}_i = \hat{b}_0 + \hat{b}_1 P_i$ (at point B).

For each data point, the regression *residual* is the difference between the actual and fitted value of the dependent variable. The residual, $\hat{e}_i$, associated with data point A in the figure, is given by $\hat{e}_i = S_i - \hat{S}_i$. The parameter values are chosen so that when all the residuals are squared and then added, the resulting sum is minimized. In this way, positive errors and negative errors are treated symmetrically; large errors are given a more-than-proportional weight.

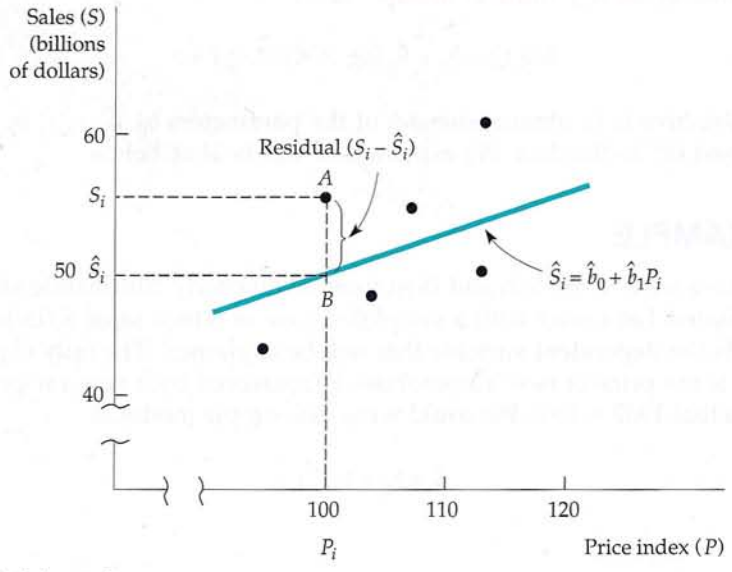

Figure A.1 Least Squares

The regression line is chosen to minimize the sum of squared residuals. The residual associated with price P_i is given by line AB.

As we will see shortly, this criterion lets us do some simple statistical tests to help interpret the regression.

As an example of estimation, let's return to the two-variable model of auto sales given by equation (A.3). The result of fitting this equation using the least-squares criterion is

$$\hat{S} = -25.5 + 0.57P \qquad \textbf{(A.5)}$$

In equation (A.5), the intercept –225.5 indicates that if the price index were zero, sales would be $ – 225.5 billion. The slope parameter indicates that a 1-unit increase in the price index for new cars leads to a $0.57 billion increase in auto sales. This rather surprising result—an upward-sloping demand curve—is inconsistent with economic theory and should make us question the validity of our model.

Let's expand the model to consider the possible effects of two additional explanatory variables: personal income I (in billions of dollars) and the rate of interest R (the three-month Treasury bill rate). The estimated regression when there are three explanatory variables is

$$\hat{S} = 51.1 - 0.42P + 0.046I - 0.84R \qquad \textbf{(A.6)}$$

The importance of including all relevant variables in the model is suggested by the change in the regression results after the income and interest rate variables are added. Note that the coefficient of the P variable has changed substantially, from 0.57 to –0.42. The coefficient –0.42 measures the effect of an increase in price on sales, *with the effect of interest rates and income held constant.* The negative price coefficient is consistent with a downward-sloping demand curve. Clearly, the failure to control for interest rates and income leads to the false conclusion that sales and price are positively related.

The income coefficient, 0.046, tells us that for every $1 billion increase in personal income in the United States, automobile sales are likely to increase by $46 million (or $0.046 billion). The interest rate coefficient reflects the fact that for every one percentage point increase in the rate of interest, automobile sales are likely to fall by $840 million. Clearly, automobile sales are very sensitive to the cost of borrowing.

STATISTICAL TESTS

Our estimates of the true (but unknown) parameters are numbers that depend on the set of observations that we started with—that is, with our **sample**. With a different sample we would obtain different estimates.[2] If we continue to collect more and more samples and generate additional estimates, the estimates of each parameter will follow a probability distribution. This distribution can be summarized by a *mean* and a measure of dispersion around that mean, a standard deviation that we refer to as the *standard error of the coefficient.*

• **sample** Set of observations for study, drawn from a larger universe.

Least-squares has several desirable properties. First, it is *unbiased.* Intuitively, this means that if we could run our regression over and over again with different samples, the average of the many estimates that we obtained for each coefficient would equal the true parameter. Second, least-squares is *consistent.* In other words, if our sample were very large, we would obtain estimates that came very close to the true parameters.

In econometric work, we often assume that the error term, and therefore the estimated parameters, are normally distributed. The normal distribution has

[2]The least-squares formula that generates these estimates is called the *least-squares estimator*, and its values vary from sample to sample.

the property that the area within 1.96 standard errors of its mean is equal to 95 percent of the total area. With this information, we can ask the following question: Can we construct an interval around $\hat{b}$ such that there is a 95-percent probability that the true parameter lies within that interval? The answer is yes, and this 95-percent *confidence interval* is given by

$$\hat{b} \pm 1.96 \text{ (standard error of } \hat{b}) \tag{A.7}$$

Thus, when working with an estimated regression equation, we must not only look at the *point* estimates but also examine the standard errors of the coefficients to determine bounds for the true parameters.[3]

If a 95-percent confidence interval contains 0, then the true parameter b may actually be zero (even if our estimate is not). This result implies that the corresponding independent variable may *not* really affect the dependent variable, even if we thought it did. We can test the hypothesis that a true parameter is actually equal to 0 by looking at its *t-statistic*, which is defined as

$$t = \frac{\hat{b}}{\text{Standard error of } \hat{b}} \tag{A.8}$$

If the t-statistic is less than 1.96 in magnitude, the 95-percent confidence interval around $\hat{b}$ must include 0. This means that we cannot reject the hypothesis that the true parameter b equals 0. We therefore say that our estimate, whatever it may be, is *not statistically significant*. Conversely, if the t-statistic is greater than 1.96 in absolute value, we reject the hypothesis that $b = 0$ and call our estimate *statistically significant*.

Equation (A.9) shows the multiple regression for the auto sales model (equation A.6) with a set of standard errors and t-statistics added:

$$\begin{array}{rccccc} \hat{S} = & 51.1 & - 0.42P & + 0.046I & - 0.84R \\ & (9.4) & (0.13) & (0.006) & (0.32) \\ t = & 5.44 & -3.23 & 7.67 & -2.63 \end{array} \tag{A.9}$$

The standard error of each estimated parameter is given in parentheses just below the estimate, and the corresponding t-statistics appear below that.

Let's begin by considering the price variable. The standard error of 0.13 is small relative to the coefficient –0.42. In fact, we can be 95 percent certain that the *true* value of the price coefficient is on the interval given by –0.42 plus or minus 1.96 standard deviations (i.e., –0.42 plus or minus [1.96][0.13] = –0.42 ± 0.25). This puts the true value of the coefficient between –0.17 and –0.67. Because this range does not include zero, the effect of price is both significantly different from zero and negative. We can also arrive at this result from the t-statistic. The t of –3.23 reported in equation (A.9) for the price variable is equal to –0.42 divided by 0.13. Because this t-statistic exceeds 1.96 in absolute value, we conclude that price is a significant determinant of auto sales.

Note that the income and interest rate variables are also significantly different from zero. The regression results tell us that an increase in income is likely to have a statistically significant positive effect on auto sales, whereas

[3]When there are fewer than 100 observations, we multiply the standard error by a number somewhat larger than 1.96.

an increase in interest rates will have a statistically significant negative effect.

GOODNESS OF FIT

Reported regression results usually contain information that tells us how closely the regression line fits the data. One statistic, the **standard error of the regression (SER)**, is an estimate of the standard deviation of the regression error term e. Whenever all the data points lie on the regression line, the SER is zero. Other things being equal, the larger the standard error of the regression, the poorer the fit of the data to the regression line. To decide whether the SER is large or small, we compare it in magnitude with the mean of the dependent variable. This comparison provides a measure of the *relative* size of the SER, a more meaningful statistic than its absolute size.

• **standard error of the regression** Estimate of the standard deviation of the regression error.

R-squared (R^2) the percentage of the variation in the dependent variable that is accounted for by all the explanatory variables, measures the overall goodness-of-fit of the multiple regression equation.[4] Its value ranges from 0 to 1. An R^2 of 0 means that the independent variables explain none of the variation of the dependent variable; an R^2 of 1 means that the independent variables explain the variation perfectly. The R^2 for the sales equation (A.9) is 0.94. This tells us that the three independent variables explain 94 percent of the variation in sales.

• **R-squared (R^2)** Percentage of the variation in the dependent variable that is accounted for by all the explanatory variables.

Note that a high R^2 does not by itself mean that the variables actually included in the model are the appropriate ones. First, the R^2 varies with the types of data being studied. Time series data with substantial upward growth usually generate much higher R^2s than do cross-section data. Second, the underlying economic theory provides a vital check. If a regression of auto sales on the price of wheat happened to yield a high R^2, we would question the model's reliability. Why? Because our theory tells us that changes in the price of wheat have little or no effect on automobile sales.

The overall reliability of a regression result depends on the formulation of the model. When studying an estimated regression, we should consider things that might make the reported results suspicious. First, have variables that should appear in the relationship been omitted? That is, is the *specification* of the equation wrong? Second, is the functional form of the equation correct? For instance, should variables be in logarithms? Third, is there another relationship that relates one of the explanatory variables (say X) to the dependent variable Y? If so, X and Y are jointly determined, and we must deal with a two-equation model, not one with a single equation. Finally, does adding or removing one or two data points result in a major change in the estimated coefficients—i.e., is the equation *robust?* If not, we should be very careful not to overstate the importance or reliability of the results.

ECONOMIC FORECASTING

A forecast is a prediction about the values of the dependent variable, given information about the explanatory variables. Often, we use regression models to generate *ex ante forecasts*, in which we predict values of the dependent variable beyond the time period over which the model has been estimated. If we know the values of the explanatory variables, the forecast is *unconditional;*

[4]The variation in Y is the sum of the squared deviations of Y from its mean. R^2 and SER provide similar information about goodness of fit, because $R^2 = 1 - \text{SER}^2/\text{Variance } (Y)$.

if they must be predicted as well, the forecast is *conditional* on these predictions. Sometimes *ex post* forecasts, in which we predict what the value of the dependent variable would have been if the values of the independent variables had been different, can be useful. An *ex post* forecast has a forecast period such that all values of the dependent and explanatory variables are known. Thus *ex post* forecasts can be checked against existing data and provide a direct means of evaluating a model.

For example, reconsider the auto sales regression discussed above. In general, the forecasted value for auto sales is given by

$$\hat{S} = \hat{b}_0 + \hat{b}_1 P + \hat{b}_2 I + \hat{b}_3 R + \hat{e} \qquad \textbf{(A.10)}$$

where $\hat{e}$ is our prediction for the error term. Without additional information, we usually take $\hat{e}$ to be zero.

Then, to calculate the forecast we use the estimated sales equation:

$$\hat{S} = 51.1 - 0.42P + 0.046I - 0.84R \qquad \textbf{(A.11)}$$

We can use (A.11) to predict sales when, for example, $P = 100$, $I = \$1$ trillion, and $R = 8$ percent. Then,

$$\hat{S} = 51.1 - 0.42(100) + 0.046(1000 \text{ billion}) - 0.84(8\%) = \$48.4 \text{ billion}$$

Note that \$48.4 billion is an *ex post* forecast for a time when $P = 100$, $I = \$1$ trillion, and $R = 8$ percent.

To determine the reliability of *ex ante* and *ex post* forecasts, we use the *standard error of forecast (SEF)*. The SEF measures the standard deviation of the forecast error within a sample in which the explanatory variables are known with certainty. Two sources of error are implicit in the SEF. The first is the error term itself, because $\hat{e}$ may not equal 0 in the forecast period. The second source arises because the estimated parameters of the regression model may not be exactly equal to the true parameters.

As an application, consider the SEF of \$7.0 billion associated with equation (A.11). If the sample size is large enough, the probability is roughly 95 percent that the predicted sales will be within 1.96 standard errors of the forecasted value. In this case, the 95-percent confidence interval is \$48.4 billion ± \$14.0 billion, i.e., from \$34.4 billion to \$62.4 billion.

Now suppose we wish to forecast automobile sales for some date in the future, such as 2007. To do so, the forecast must be conditional because we need to predict the values for the independent variables before calculating the forecast for automobile sales. Assume, for example, that our predictions of these variables are as follows: $\hat{P} = 200$, $\hat{I} = \$5$ trillion, and $\hat{R} = 10$ percent. Then, the forecast is given by $\hat{P} = 51.1 - 0.42(200) + 0.046(5000 \text{ billion}) - 0.84(10) = \$188.7$ billion. Here \$188.7 billion is an *ex ante* conditional forecast.

Because we are predicting the future, and because the explanatory variables do not lie close to the means of the variables throughout our period of study, the SEF is equal to \$8.2 billion, which is somewhat greater than the SEF that we calculated previously.[5] The 95-percent confidence interval associated with our forecast is the interval from \$172.3 billion to \$205.1 billion.

[5]For more on SEF, see Pindyck and Rubinfeld, *Econometric Models and Economic Forecasts*, ch. 8.

EXAMPLE A.1 The Demand for Coal

Suppose we want to estimate the demand for bituminous coal (given by sales in tons per year, COAL) and then use the relationship to forecast future coal sales. We would expect the quantity demanded to depend on the price of coal (given by the Producer Price Index for coal, PCOAL) and on the price of a close substitute for coal (given by the Producer Price Index for natural gas, PGAS). Because coal is used to produce steel and electricity, we would also expect the level of steel production (given by the Federal Reserve Board Index of iron and steel production, FIS) and electricity production (given by the Federal Reserve Board Index of electric utility production, FEU) to be important demand determinants.

Our model of coal demand is therefore given by the following equation:

$$\text{COAL} = b_0 + b_1\,\text{PCOAL} + b_2\,\text{PGAS} + b_3\,\text{FIS} + b_4\,\text{FEU} + e$$

From our theory, we would expect b_1 to be negative because the demand curve for coal is downward sloping. We would also expect b_2 to be positive because a higher price of natural gas should lead industrial consumers of energy to substitute coal for natural gas. Finally, we would expect both b_3 and b_4 to be positive because the greater the production of steel and electricity, the greater the demand for coal.

This model was estimated using monthly time-series data covering eight years. The results (with t-statistics in parentheses) are

$$\begin{array}{lccccc} \text{COAL} = 12{,}262 & + 92.34\,\text{FIS} & + 118.57\,\text{FEU} & - 48.90\,\text{PCOAL} & + 118.91\,\text{PGAS} \\ \quad (3.51) & \quad (6.46) & \quad (7.14) & \quad (-3.82) & \quad (3.18) \end{array}$$

$$R^2 = 0.692 \qquad \text{SER} = 120{,}000$$

All the estimated coefficients have the signs that economic theory would predict. Each coefficient is also statistically significantly different from zero because the t-statistics are all greater than 1.96 in absolute value. The R^2 of 0.692 says that the model explains more than two-thirds of the variation in coal sales. The standard error of the regression SER is equal to 120,000 tons of coal. Because the mean level of coal production was 3.9 million tons, SER represents approximately 3 percent of the mean value of the dependent variable. This suggests a reasonably good model fit.

Now suppose we want to use the estimated coal demand equation to forecast coal sales up to one year into the future. To do so, we substitute values for each of the explanatory variables for the 12-month forecasting period into the estimated equation. We also estimate the standard error of forecast (the estimate is 0.17 million tons) and use it to calculate 95-percent confidence intervals for the forecasted values of coal demand. Some representative forecasts and confidence intervals are given in Table A.1.

TABLE A.1 Forecasting Coal Demand

	Forecast	Confidence Interval
1-month forecast (tons)	5.2 million	4.9–5.5 million
6-month forecast (tons)	4.7 million	4.4–5.0 million
12-month forecast (tons)	5.0 million	4.7–5.3 million

SUMMARY

1. Multiple regression is a statistical procedure for quantifying economic relationships and testing hypotheses about them.
2. The linear regression model, which relates one dependent variable to one or more independent variables, is usually estimated by choosing the intercept and slope parameters that minimize the sum of the squared residuals between the actual and predicted values of the dependent variable.
3. In a multiple-regression model, each slope coefficient measures the effect on the dependent variable of a change in the corresponding independent variable, holding the effects of all other independent variables constant.
4. A t-test can be used to test the hypothesis that a particular slope coefficient is different from zero.
5. The overall fit of the regression equation can be evaluated using the standard error of the regression (SER) (a value close to zero means a good fit) or R^2 (a value close to one means a good fit).
6. Regression models can be used to forecast future values of the dependent variable. The standard error of forecast (SEF) measures the accuracy of the forecast.