Statistical Models for the Social and Behavioral Sciences

Multiple Regression and Limited-Dependent Variable Models

WILLIAM H. CROWN
CONCEPTUAL BASIS FOR REGRESSION ANALYSIS

Regression analysis is a technique that enables researchers to test hypotheses about relationships between variables in a much more powerful way than is possible with the bivariate methods used in the previous chapter. Regression analysis predicts the value for a "dependent" variable given the values for one or more "explanatory" or "independent" variables. The use of regression analysis, however, requires having a theory about the nature of the relationships among the variables. Policy analysts and social scientists often work with samples containing a thousand or more variables and tens of thousands of cases or observations. Researchers wading into data bases of this magnitude without the aid of a theory soon find themselves awash in printouts and hopelessly confused.

Figure 3.1 illustrates the basic concept of regression analysis for the simplest case, involving only one explanatory variable. In the case of one explanatory variable two regression coefficients must be estimated. The constant term of the regression line, \( b_0 \), measures the expected value for the dependent variable when the value of the explanatory variable is zero. The slope coefficient, \( b_1 \), measures the change in the dependent variable resulting from a one-unit increase in the explanatory variable. Regression analysis estimates the line that best describes the linear association between the dependent variable \( Y \) and the explanatory variable \( X \). This line is completely described by the two coefficients \( b_0 \) and \( b_1 \).

DERIVATION OF THE REGRESSION MODEL

Assume for the moment that we have somehow estimated the line in Figure 3.1. This line enables us to predict a \( Y \) value corresponding to each \( X \) value. However, the predictions will seldom be perfect. Sometimes the regression line will underpredict the actual value of \( Y \) given a particular value for \( X \); other times, it
Figure 3.1
Illustration of Simple Regression

\[ \hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 \]

will overpredict \( Y \). The differences between the actual values for \( Y \) and the predicted values \( \hat{Y} \) are known as the errors or residuals \( \epsilon \).

It would clearly be desirable if regression analysis found values for \( \hat{b}_0 \) and \( \hat{b}_1 \) that minimize these errors in some way. For example, one possibility would be to find the values of \( \hat{b}_0 \) and \( \hat{b}_1 \) that minimize the sum of the errors around the regression line. But if one uses this approach, any regression line passing through the middle of the scatterplot would produce positive and negative errors that would exactly cancel each other out. There are an infinite number of such lines, so minimizing the sum of the errors is not a very helpful criterion in estimating values for \( \hat{b}_0 \) and \( \hat{b}_1 \).

An alternative criterion would be to find the values for \( \hat{b}_0 \) and \( \hat{b}_1 \) that minimize the sum of the absolute values of the errors. This solves the problem of the positive and negative errors canceling each other out but it is difficult mathematically to find the values of \( \hat{b}_0 \) and \( \hat{b}_1 \) that minimize the sum of the absolute errors. This is because calculus cannot be used to solve for the maximum and minimum points of absolute value functions.

On the other hand, it is a straightforward matter to find the values of \( \hat{b}_0 \) and \( \hat{b}_1 \) that minimize the sum of the squared errors around the regression line. Squaring the errors has several advantages. First, the negative errors squared no longer cancel the positive errors, so they contribute to the total sum of squared errors as desired. Second, squaring the errors increases the effect of large errors on the slope coefficients of the regression line, this highlights outliers or extreme cases. Finally, squaring the errors enables differential calculus to be used for finding the coefficient values that minimize the sum of squared errors. (For interested readers, a very brief summary of differential calculus is provided in Appendix A.)
Recall that the regression residuals or errors are equal to the difference between the actual and predicted values for Y:

\[ \hat{e}_i = (Y_i - \hat{Y}_i) = (Y_i - \hat{b}_0 - \hat{b}_1 X_i) \]

Squaring these errors and summing over all of the observations results in the sum of the squared errors.

\[ \sum e_i^2 = \sum (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2 \]  

(1)

The values for \( \hat{b}_0 \) and \( \hat{b}_1 \) that minimize equation (1) can be found by taking the partial derivative of (1) with respect to \( \hat{b}_0 \) and \( \hat{b}_1 \) and setting the results equal to zero:

\[ \frac{\partial}{\partial \hat{b}_0} \sum e_i^2 = -2 \sum (Y_i - \hat{b}_0 - \hat{b}_1 X_i) = 0 \]

\[ \frac{\partial}{\partial \hat{b}_1} \sum e_i^2 = -2 \sum (Y_i - \hat{b}_0 - \hat{b}_1 X_i) X_i = 0 \]

This results in two equations with two unknown values, \( \hat{b}_0 \) and \( \hat{b}_1 \). These equations can be solved for the computational formulas that will yield estimates of \( \hat{b}_0 \) and \( \hat{b}_1 \):

\[ \hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X} \]  

(2)

\[ \hat{b}_1 = \frac{\sum X_i Y_i - \bar{X} \bar{Y} \sum X_i}{\sum X_i^2 - N \bar{X}^2} \]  

(3)

Since \( \hat{b}_0 \) and \( \hat{b}_1 \) are estimated with sample data, they have sampling distributions just like any other sample statistic (such as \( \bar{X} \)). That is, given 1,000 different samples, it would be possible to estimate \( \hat{b}_0 \) and \( \hat{b}_1 \) for each of these samples. The result would be sampling distributions for \( \hat{b}_0 \) and \( \hat{b}_1 \). As with inferential statistics, the statistical significance of \( \hat{b}_0 \) and \( \hat{b}_1 \) depends upon the sizes of these coefficients relative to the standard errors of their respective sampling distributions. The formulas for the standard errors of \( \hat{b}_0 \) and \( \hat{b}_1 \) are tedious to derive without matrix algebra, but it can be shown that they are given by the following:

\[ s_{\hat{b}_0} = \sqrt{\frac{\sum e_i^2}{N - 2}} \frac{\sum X_i^2}{N \sum (X_i - \bar{X})^2} \]  

(4)
\[ S_b^2 = \frac{\sum \hat{e}_i^2}{(N-2)\sum (X_i - \bar{X})^2} \]  

Note that the standard errors for \( \hat{b}_0 \) and \( \hat{b}_1 \) are both a function of the sum of squared errors for the regression.

**STATISTICAL SIGNIFICANCE OF COEFFICIENT ESTIMATES**

Given a set of estimates for the regression coefficients and their standard errors, the null hypotheses that \( b_0 \) equals a particular value \( b_0^* \) and that \( b_1 \) equals \( b_1^* \) can be tested with:

\[ t_{b_0} = \frac{\hat{b}_0 - b_0^*}{s_{\hat{b}_0}} \]  
\[ t_{b_1} = \frac{\hat{b}_1 - b_1^*}{s_{\hat{b}_1}} \]  

Equations (6) and (7) each have a \( t \)-distribution with \( N - k + 1 \) degrees of freedom, where \( N \) is the sample size and \( k \) is the number of explanatory variables in the regression. In a regression that includes a constant term and one explanatory variable, \( k + 1 \) is equal to 2.

Researchers are often more interested in the signs and statistical significance of the estimated coefficients than in their specific numerical values. The reasons for this will be discussed in detail later. For now, it is sufficient to note that the values of the estimated coefficients are frequently sensitive to the other variables included in (or excluded from) the model, as well as the functional form used in specifying the model. In any event, when researchers wish to test the statistical significance of an estimated coefficient, it is common practice to test whether the coefficient is significantly different from zero. This is because a zero slope coefficient implies that there is no linear relationship between the explanatory variable and the dependent variable.

Thus, the usual \( t \)-tests for the statistical significance of estimated coefficients in the simple regression model are:

\[ t_{b_0} = \frac{\hat{b}_0}{s_{\hat{b}_0}} \]  
\[ t_{b_1} = \frac{\hat{b}_1}{s_{\hat{b}_1}} \]  

To determine if the calculated \( t \)-ratios are statistically significant, they are compared to the tabulated value of \( t \) for the selected alpha level (e.g., 0.05); if the calculated value exceeds the tabulated value, the coefficient is considered to be
statistically significant. Note that it is perfectly possible (and indeed likely in larger
tests) that some coefficients will be statistically significant and others will not.

If there is no expectation about the sign of the coefficient, the researcher
can replace the calculated value with the tabulated value for $t$, assuming one-half of the
alpha region in each tail. The $p$-values produced by most statistical packages
assumes such two-tailed tests. On the other hand, it is often the case that the
researcher will have a theoretical expectation about the sign of an estimated
coefficient. If this is the case, the entire alpha region should be concentrated in one
tail. Concentrating alpha region in one tail lowers the critical value of the $t$-statistic
that must be exceeded in order to reject the null hypothesis that the coefficient is
significantly different from zero. For example, in a very large sample, the critical
$t$-value for a two-tailed test at the 95 percent confidence level is 1.96; for the one-
tailed test it is 1.67.

**GOODNESS-OF-FIT MEASURES**

At this point in the discussion, formulas have been presented for obtaining
numerical estimates of the regression coefficients, as well as tests of the statistical
significance of these coefficients. It would be useful, however, to also have a
measure for how well the model as a whole describes the variation in the dependent
variable.

The total deviation of a particular point from its mean is given by \((Y_i - \bar{Y})\). The
difference between the observed value for \(Y_i\) and the predicted value from the
regression is given by \((Y_i - \hat{Y}_i)\). Because this is the error for a single sample point,
it is reasonable to consider this to be the portion of the total deviation that is
not explained by the model. The remaining portion of the total deviation \((\bar{Y} - \bar{Y})\) is,
therefore, the portion that is explained by the model. Thus, the total deviation of a
particular observation of the dependent variable from its mean can be
decomposed into two pieces—the piece explained by the regression model and the
piece that is not explained by the model:

\[(Y_i - \bar{Y}) = (\bar{Y} - \bar{Y}) + (Y_i - \hat{Y}_i)\]  

(8)

Squaring each deviation and summing over all values of the dependent variable
yields an expression for the total variation of \(Y\) around its mean:

\[\sum (Y_i - \bar{Y})^2 = \sum (\bar{Y} - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2\sum (\bar{Y} - \bar{Y})(Y_i - \hat{Y}_i)\]  

(9)

Note that the terms \(2\sum (\bar{Y} - \bar{Y})(Y_i - \hat{Y}_i)\) arise because of the cross products between
\((Y_i, \bar{Y})\) and \((Y_i, \hat{Y}_i)\) that occur when one squares the right-hand side of equation (8).

It was mentioned earlier that regression analysis is based on many
assumptions. One of these assumptions is that the expected value of the errors is zero. This,
in turn, implies that the sum of the cross product terms in equation (9) is zero.
Therefore, the total variation of the dependent variable around its mean is equal to
the unexplained variation plus the explained variation:
\[ \sum (y_i - \hat{y}_i)^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (\bar{y} - \bar{y})^2 \]  

(10)

Phrased somewhat differently, the total sum of squares (TSS) of the dependent variable is equal to the regression sum of squares (RSS) plus the sum of squared errors (SSE):

\[ \text{TSS} = \text{RSS} + \text{SSE} \]  

(11)

Dividing both sides of equation (11) by TSS yields:

\[ 1 = \frac{\text{RSS}}{\text{TSS}} + \frac{\text{SSE}}{\text{TSS}} \]

This expresses the TSS in terms of two proportions—the proportion explained by the regression and the proportion that is not explained by the regression.

The proportion of the total variation of \( Y \) around its mean that is explained by the regression model is known as \( R^2 \) and is found by:

\[ R^2 = \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\text{SSE}}{\text{TSS}} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \]  

(12)

\( R^2 \) varies between 0 and 1. High values for \( R^2 \) indicate that the model explains a high proportion of the total variation in \( Y \) around its mean. For example, a value of 0.90 would mean that the regression model explained 90 percent of the total variation in \( Y \).

Intuitively, it might seem that high \( R^2 \) values would be indicative of a good model. Yet this is not always true. For one thing, \( R^2 \) values tend to be higher for some types of data than for others. Regression models estimated using time-series data (data collected on variables over time) tend to have higher \( R^2 \)'s than regression models estimated using cross-sectional data. For example, a time-series regression model of a macroeconomic variable, such as the Gross National Product, could easily have an \( R^2 \) value greater than 90 percent. In contrast, a cross-sectional model of the determinants of personal earnings might have an \( R^2 \) of only 20 percent.

Correlation values for regression models also tend to vary systematically with the aggregation level of the data—models based on aggregate data tend to have higher \( R^2 \)'s than models based on micro data (data for individuals, families, households, etc.). Of course, the level of data aggregation is often related to whether the data are time-series or cross-sectional in nature. Macroeconomic time-series data are also aggregate data. In contrast, individual-level earnings data are often measured with cross-sectional surveys such as the Decennial Census or the Current Population Survey. But there are many exceptions. Longitudinal studies of individuals, households, or families (such as the Retirement History Survey or...
AN EXAMPLE OF SIMPLE REGRESSION

Suppose that you are a federal budget analyst and your job is to figure out what is creating the persistent federal deficits. Like everyone else in Washington, you read the newspapers which seem to be saying that the elderly are "busting" the budget. To see if this is the case, you collect information on the size of the population age 65 and over and on the federal budget deficit over the years from 1976 to 1985. This information is displayed in the first two columns of Table 3.1.

The first task in estimating the equation is to calculate the coefficients of the regression line. Recall from equation (3) that

\[ \hat{\beta}_1 = \frac{\sum X_i Y_i - N \bar{X} \bar{Y}}{\sum X_i^2 - N\bar{X}^2} \]

where \( Y \) are the values for the dependent variable (federal deficits) and \( X \) are the values for the explanatory variable (elderly population size). To use this formula to estimate the slope coefficient \( \hat{\beta}_1 \) it is first necessary to calculate the mean of the dependent variable \( \bar{Y} \), the sum of the squared values of \( X \), and the sum of the product of each \( X \) value times each \( Y \) value, \( \sum X_i Y_i \). \( N \) refers to the sample size—in this case, 10.

The calculation of \( \bar{Y} \) and \( \bar{X} \) is carried out in the usual fashion and is shown at the bottom of columns 1 and 2 in Table 3.1. The calculation of \( \sum X_i Y_i \) is shown in column 3 of Table 3.1. The entries in column 3 are found by multiplying the corresponding \( Y \) and \( X \) values in columns 1 and 2. Summing these products results in \( \sum X_i Y_i \) shown at the bottom of column 3. Finally, to generate the \( \sum X_i^2 \), the \( X \) values in column 2 are squared and all of these squared values are summed. The results are shown in column 4 of Table 3.1. Substituting all of these values into equation (3) yields:

\[ \hat{\beta}_1 = \frac{29781.9 - (10)(25.94)(111.27)}{6756.54 - (10)(25.94)^2} \]

\[ = 33.16 \]

This coefficient indicates that for each one-million-person increase in the population age 65 and over, the federal deficit increases by 33.16 billion dollars.
<table>
<thead>
<tr>
<th>Year (in Millions)</th>
<th>Column (a)</th>
<th>Column (b)</th>
<th>Column (c)</th>
<th>Column (d)</th>
<th>Column (e)</th>
<th>Column (f)</th>
<th>Column (g)</th>
<th>Column (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Calculation for Simple Regression of Federal Deficits
Once the slope coefficient $\hat{b}_1$ has been estimated, the estimation of the constant term is easy. From equation (2) the coefficient for the constant term is:

$$\hat{b}_0 = 111.27 - (33.16)(25.94)$$

$$= -748.90$$

The estimated value for $\hat{b}_0$ means that if the elderly population were equal to zero, we would expect the federal deficit to be -748.90 billion dollars— in other words, a surplus of $748.9$ billion!

Of course, these estimates will have no policy importance unless they are statistically significant. To test if the coefficients are statistically different from zero, the coefficient values are divided by their standard errors. The resulting ratios are $t$-statistics that can be used for hypotheses testing.

To calculate the standard errors for $\hat{b}_0$ and $\hat{b}_1$, two additional pieces of information are needed—the sum of the squared errors from the regression equation $\sum e_i^2$ and the sum of squared deviations of the explanatory variable (elderly population) from its mean, $\sum (X_i - \bar{X})^2$.

Calculation of $\sum e_i^2$ is shown in columns 5 and 6 of Table 3.1. To estimate the residuals from the regression equation, the predicted values for the dependent variable $\hat{Y}$, generated by the model are first calculated. These are shown in column 5.

To calculate the predicted values, each of the $X_i$ values is substituted into the estimated equation. For example, in 1976 the elderly population was 23.3 million. The value 23.3 is multiplied by the slope coefficient 33.16, and the result is added to the constant of -748.9. This yields a $\hat{Y}$ estimate for 1976 of 23.72:

$$23.72 = -748.9 + 33.16(23.3)$$

In this manner, a predicted value for $\hat{Y}$ is generated for each year in the sample. The errors shown in column 6 of Table 3.1 are calculated as the difference between the actual value of $Y$ (column 1) and the predicted value (column 5). Squaring each of the errors in column 6 yields the squared errors in column 7. Finally, summing the values in column 7 yields the sum of the squared errors, $\sum e_i^2$.

The squared deviations of the explanatory variable around its mean, shown in column 8 of Table 3.1, are found simply by subtracting the mean elderly population for the sample as a whole from the elderly population in each year and squaring the results. Summing all of the entries in column 8 yields the sum of the squared deviations, $\sum (X_i - \bar{X})^2$.

It is now possible to calculate the standard errors for $\hat{b}_0$ and $\hat{b}_1$. Substituting the appropriate values in the equations for the standard errors of $\hat{b}_0$ and $\hat{b}_1$ yields
The \(t\)-statistics can now be calculated by taking the ratio of each coefficient to its standard error:

\[
\frac{b_0}{s_{b_0}} = \frac{-748.9}{172.28} = -4.35
\]

\[
\frac{b_1}{s_{b_1}} = \frac{33.16}{6.63} = 5.00
\]

Both of these \(t\)-statistics have 8 degrees of freedom \((N-2)\). One degree of freedom is used up in estimating the slope coefficient; another is used up in estimating the constant term. At a 95 percent confidence level, the tabulated \(t\)-value for 8 degrees of freedom is 2.306. The calculated \(t\)-statistics for both coefficients are greater in absolute value than the critical value of the \(t\)-statistic from the table. This means
that the null hypotheses that the constant term and slope coefficient are equal to zero are rejected.

Thus far, we have calculated the coefficients of the regression model and determined that they are statistically different from zero. But we still do not have a sense of how much of the variation in federal budget deficits is explained by elderly population size. This measure is provided by the $R^2$ of the model.

From equation (12) it can be seen that the $R^2$ is a function of the sum of the squared errors from the regression $\sum \hat{e}_i^2$ and the sum of the squared deviations of $Y$ around its mean $\sum (Y_i - \bar{Y})^2$. Calculation of $\sum (Y_i - \bar{Y})^2$ is shown in column 9 of Table 3.1.

Substituting into equation (12), we can calculate the $R^2$ as follows:

$$R^2 = 1 - \frac{\sum \hat{e}_i^2}{\sum (Y_i - \bar{Y})^2}$$

$$= 1 - \frac{9734.49}{40233.47}$$

$$= .76$$

Thus, according to the model, elderly population size explains 76 percent of the variation in federal budget deficits.

In short, the model indicates that elderly population size does a very respectable job of predicting the size of the federal budget deficit. The $t$-statistic for elderly population size is highly significant and the model explains 76 percent of the variation in the federal budget deficit. There are, however, at least three major reasons why we should be dubious about the conclusion that elderly population size is such a significant predictor of the federal deficit.

First, the use of data collected over time (time series) will tend to generate models with high $R^2$'s and inflated $t$-statistics—a problem known as autocorrelation. Autocorrelation is discussed briefly later in this chapter and, in more detail, in Chapter 5.

Second, the $R^2$ of the model is probably overstated because of the aggregate nature of the data. As noted earlier, regression models based on highly aggregated data tend to have higher $R^2$'s than models based on less highly aggregated data. In the current example, the high aggregation level of the data interacts with the autocorrelation problem to overstate the $R^2$.

A third major problem with the simple model discussed above is that it contains only one explanatory variable. It is possible that the size of the older population does contribute to the size of federal deficits, but it is likely that other factors such as unemployment, wars, and the size of the child population also contribute to federal spending. Moreover, to properly model deficits, it is necessary to include not only variables that affect federal spending, but also variables that affect tax revenues such as income levels. It is likely that, after we account for all of these
MULTIPLE REGRESSION

Extending the model to include two or more explanatory variables is straightforward. The sum of squared errors (SSE) from a multiple regression equation with K variables is minimized by taking the partial derivative of the SSE expression with respect to each variable and the constant term:

\[ SSE = \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{12} - \ldots - \hat{\beta}_K X_{1K})^2 \]

\[ \frac{\partial SSE}{\partial \hat{\beta}_0} = -2 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{12} - \ldots - \hat{\beta}_K X_{1K}) \]

\[ \frac{\partial SSE}{\partial \hat{\beta}_1} = -2 \sum (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{12} - \ldots - \hat{\beta}_K X_{1K}) X_{1i} \]

This results in K+1 equations with K+1 unknown coefficients. If the derivatives are set equal to zero, these K+1 equations can be solved to generate computational formulas for each of the coefficients in terms of the sample variables. Unfortunately, the formulas for the slope coefficients become very unwieldy for models with two or more explanatory variables unless matrix algebra is used. This is why, in the interest of generalizability, advanced statistics texts make heavy use of matrix algebra. In matrix notation, the formula for the constant term and slope coefficients is always:

\[ B = (X'X)^{-1}X'Y \]

where B is a column vector of coefficients, X is a table or matrix of data for the explanatory variables, and Y is a column vector containing the observations for the dependent variable. The beauty of matrix notation is that this formula is always the same, whether the model has 1 explanatory variable or 30. Moreover, the same generalizability pertains to other results, such as the formulas for standard errors.

As in the case of simple regression, the constant term in a multiple regression model represents the expected value for the dependent variable when all of the explanatory variables are equal to zero. The slope coefficient for each variable represents the change in the dependent variable with respect to a one-unit increase in that particular explanatory variable controlling for the effects of the other variables included in the equation.
Also as in the case of simple regression, the statistical significance of each variable in a multiple regression can be assessed with a t-test based on the ratio of each estimated coefficient to its standard error. In multiple regression, it is very possible, indeed likely, that some variables will be statistically significant and others will not. In a model with \( K \) variables, the t-statistics have \( N - (K + 1) \) degrees of freedom.

**BETA COEFFICIENTS**

Having estimated coefficients for several variables, one is often tempted to compare the magnitudes of the coefficients as a measure of the relative strength of the relationship between particular explanatory variables and the dependent variable. This is not valid because the sizes of the coefficients for the different explanatory variables are a function of the units that the explanatory variables are measured in. To be able to compare the relative size of the impacts of different explanatory variables on the dependent variable, it is necessary to first standardize the units that these variables are measured in.

One way to do this is to construct a new set of variables that are expressed in terms of standard deviations from their means:

\[
\frac{(Y_i - \bar{Y})}{S_Y} = B_0 \frac{(X_{i1} - \bar{X}_{1})}{S_{X_1}} + B_1 \frac{(X_{i2} - \bar{X}_{2})}{S_{X_2}} + \ldots + B_k \frac{(X_{ik} - \bar{X}_{k})}{S_{X_k}}
\]

These new variables are z-scores corresponding to the original variables. Using these z-scores, it is possible to estimate a regression model where the slope coefficients represent the standard deviation change in the dependent variable associated with a standard deviation increase in an explanatory variable, controlling for all of the other variables in the equation. These slope coefficients are known as beta coefficients to distinguish them from ordinary regression coefficients. Note that no constant term is included in this equation because it makes no sense to talk about an expected standard deviation change in the dependent variable when the variation in all of the explanatory variables is set to zero.

It turns out that there is an easier way to estimate beta coefficients than the method just described. Beta coefficients can be easily calculated by multiplying each ordinary slope coefficient of the regression model by the ratio of the standard deviation of that variable to the standard deviation of the dependent variable:

\[
\hat{B}_i = B_i \frac{S_{X_i}}{S_Y}
\]

Easier still, most statistical packages produce beta coefficients as part of the standard multiple regression output or offer them as an option. Because beta coefficients measure standard deviation changes, the beta coefficients for different variables can be directly compared. The variable with the largest absolute beta coefficient (either positive or negative) has the greatest impact on the dependent variable.
variable; the variable with the next largest beta coefficient has the next largest impact; and so on.

It is usually the case that the largest beta coefficients in an equation also have the highest $t$-statistics. But this need not be the case. The beta coefficients are just the regression coefficients multiplied by the standard deviations of the explanatory variables and divided by the standard deviation of the dependent variable. If estimated regression coefficients are large, and/or the standard deviations of the explanatory variables are large relative to the dependent variable, beta coefficients can be large even if the corresponding explanatory variables are not statistically significant. The statistical significance of beta coefficients is the same as that of the regression coefficients upon which they are based. If the regression coefficients are not statistically significant, neither are the beta coefficients — no matter how large they are.

**ADJUSTED $R^2$**

In the discussion of the simple regression model, it was noted that the $R^2$ measure is sensitive to the aggregation level of the data, whether the data are time-series or cross-sectional, and other factors. Another difficulty with the $R^2$ measure is that its value approaches one as the number of variables in the regression approaches the sample size — irrespective of the explanatory power of the variables. An alternative measure, known as the adjusted $R^2$ (or $R^2_A$), corrects for this tendency by assessing a degrees-of-freedom penalty to the $R^2$ measure for each explanatory variable included in the model:

$$R^2_A = 1 - \frac{(1-R^2)(N-1)}{N-k}$$

(13)

The addition of a variable to a regression equation only increases the adjusted $R^2$ if the variable's explanatory power is greater than the degrees-of-freedom correction. Note from equation (13) that the adjusted $R^2$ equals $R^2$ in the case of simple regression. In models involving more than one explanatory variable, however, the adjusted $R^2$ will always be less than the unadjusted $R^2$. In fact, the adjusted $R^2$ can be negative! In this last instance, the researcher would do well to rethink the theoretical model underlying the regression or reexamine the data used to estimate the model. Sometimes, when an estimated model performs so poorly, it is the result of a data-processing problem such as incorrectly reading variables from a dataset. Data quality issues can be assessed by examining the descriptive statistics for all variables before the model is estimated.

It should be clear from the preceding discussion that the model with the highest $R^2$, or even the highest adjusted $R^2$, is not necessarily the best. The signs and statistical significance of the independent variables are much more important criteria for evaluating a model than the proportion of variation in the dependent variable that the model explains. We have already discussed the individual $t$-tests of significance for each variable. A more general test of the model is to evaluate
the null hypothesis that all of the slope coefficients of the explanatory variables are jointly equal to zero:

\[ H_0: b_1 = b_2 = b_3 = \ldots = b_k = 0 \]

\[ H_1: b_1 \neq b_2 \neq b_3 \neq \ldots \neq b_k \neq 0 \]

This null hypothesis can be evaluated by using the F-statistic. As one might expect, the F-statistic is a function of the model's \( R^2 \):

\[
F_{k, N-(k+1)} = \frac{R^2}{1 - R^2} \left( \frac{N-(k+1)}{k} \right)
\]

If none of the slope coefficients are statistically different from zero then there is no systematic variation of the dependent variable with any of the explanatory variables. In this instance, the F-statistic will indicate that the model is not statistically significant. A model with no significant variables will also tend to have a low \( R^2 \) (except in the case where the number of variables approaches the sample size). Consequently, models with low \( R^2 \)'s tend to have low F-statistics. Conversely, models with high \( R^2 \)'s tend to have large F-statistics. The magnitude of the F-statistic needed to reject the null hypothesis that all of the slope coefficients are jointly equal to zero varies with the number of variables in the equation, the number of observations in the sample, and the level of statistical significance at which the test is carried out. Moreover, if just one variable is statistically significant, the null hypothesis that all of the slope coefficients are equal to zero is rejected. Usually, at least one variable will be statistically significant so the standard F-test for the significance of a regression is not very interesting—it almost always indicates that the model is statistically significant.

A MULTIPLE REGRESSION EXAMPLE

Table 3.2 reports the multiple regression results for a very simple model of hours worked as a function of a person's education level, age, wages, and other income. Regression coefficients must be carefully interpreted because they are sensitive to the units that the dependent and independent variables are measured in. For example, it is not safe to assume that years of education have the biggest effect on hours worked just because this variable has the largest slope coefficient. To properly interpret the coefficient magnitudes, it is necessary to take account of how the variables are measured. In Table 3.2, the dependent variable is the number of hours per week an individual usually worked in 1989. Education level and age are measured in years, wages are measured in dollars per hour, and other income is measured in dollars per year. Thus, a one-unit (one-year) increase in education level would increase average hours worked by 0.51 hours (or about 30 minutes). Similarly, a one-year increase in age would increase individual labor supply by 0.09 hours; a one-dollar-per-hour increase in wages would increase labor supply by 0.12 hours; and each dollar of additional other income would decrease hours
Table 3.2
A Multiple Regression Model of Individual Hours Worked

<table>
<thead>
<tr>
<th>Variable</th>
<th>Regression Coefficient</th>
<th>Standard Error</th>
<th>t</th>
<th>Beta Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>0.5090</td>
<td>0.0772</td>
<td>6.590</td>
<td>0.1095</td>
</tr>
<tr>
<td>Age</td>
<td>0.087</td>
<td>0.0161</td>
<td>5.505</td>
<td>0.0954</td>
</tr>
<tr>
<td>Wages</td>
<td>0.1218</td>
<td>0.0259</td>
<td>4.708</td>
<td>0.0795</td>
</tr>
<tr>
<td>Other Income</td>
<td>-0.0004</td>
<td>0.00004</td>
<td>-9.461</td>
<td>-0.1613</td>
</tr>
<tr>
<td>Constant</td>
<td>27.5372</td>
<td>1.1915</td>
<td>23.112</td>
<td></td>
</tr>
</tbody>
</table>

$R^2 = 0.042$
$R^2 = 0.041$
$F_{cal} = 42.1618$

supplied by 0.0004 hours. Setting the effects of education, age, wages, and other income to zero, we would expect individuals to supply an average of 27.5 hours of work per week.

The second column in Table 3.2 reports the standard errors for the coefficients in the model. Recall that the standard error for each coefficient is the hypothetical standard deviation of the sampling distribution calculated from a large number of parameter estimates based upon many different samples of a fixed size. Some of these samples would generate relatively low parameter estimates; others would generate relatively high estimates. Most would be somewhere in between.

The estimated regression coefficients divided by their standard errors will have $t$-distributions. Except for very small sample sizes the $t$-distribution is a close approximation of the normal distribution. This suggests that $t$-values greater than 1.96 will usually be statistically significant at the 95 percent confidence level, assuming a two-tailed test. Put somewhat differently, when the $t$-value is greater than 1.96, we can be 95 percent sure that the coefficient is not equal to zero. This implies that there is a relationship between the explanatory variable and the dependent variable. Based on this rule of thumb, it is apparent that education, age, hourly wage, and other income are all statistically significant predictors of individual labor supply.

Note, in this example, that all of the explanatory variables are statistically significant. This need not be the case, however. Most models will contain a mixture of variables—some of which will have coefficients significantly different from zero and some of which will not. In such cases, researchers will sometimes make the error of saying that the variable "had the expected sign but was not statistically significant." Such an inference really makes no sense because the fact that the coefficient is statistically insignificant at a given confidence level means that the null hypothesis that the coefficient is equal to zero cannot be rejected.
The final column in Table 3.2 reports the beta coefficients for the model. These are the coefficients that one would get if the dependent and explanatory variables were all converted to z-scores prior to running the regression and the regression was then estimated without a constant term. The advantage of the beta coefficients relative to the regular regression coefficients is that the beta coefficients are unitless—they represent the standard deviation change in the dependent variable with respect to a one-standard-deviation increase in an explanatory variable, controlling for the other variables in the model. Beta coefficients can be directly compared. The largest beta coefficient (in absolute value) has the biggest effect on the variation of the dependent variable, the next largest beta coefficient has the second biggest effect, and so on.

Most statistical packages produce the beta coefficients as part of the standard output or offer them as an option. In the event that a particular package does not offer beta coefficients as part of the output, however, they are easy to calculate. One simply takes the estimated regression coefficient, multiplies it by the standard deviation (not the standard error) of the explanatory variable, and divides the result by the standard deviation of the dependent variable.

In the current example, the standard deviation of the dependent variable was 12.81; it was 2.76 for education level, 13.79 for age, 8.36 for wages, and 4905.56 for other income. Using these standard deviations, in conjunction with the regression coefficients in Table 3.2, we can find the beta coefficients for the explanatory variables as follows:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Beta Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>0.1095</td>
<td>0.5900</td>
</tr>
<tr>
<td>Age</td>
<td>0.0954</td>
<td>0.0887</td>
</tr>
<tr>
<td>Wages</td>
<td>0.0795</td>
<td>0.1218</td>
</tr>
<tr>
<td>Other Income</td>
<td>-0.1613</td>
<td>-0.00042</td>
</tr>
</tbody>
</table>

Table 3.2 reports three measures of the overall goodness of fit of the model. The $R^2$ value of 0.042 indicates that the model explains 4.2 percent of the variation of the dependent variable around its mean. As discussed earlier in this chapter, however, the $R^2$ measure can be a misleading indicator of goodness of fit. Its value will approach 1.0 (indicating a perfect fit) as the number of explanatory variables approaches the sample size—even if none of the explanatory variables are statistically significant! This situation tends to occur when working with aggregate data, such as national macroeconomic statistics or state-level data, where the number of observations is limited.

On the other hand, micro data on individuals, families, or households tend to have low $R^2$s because there is so much variation in individual behavior. Low $R^2$s
do not necessarily mean that the model is poor. This is because micro level models tend to produce more efficient parameter estimates than models based on aggregate data.

In either event, it is best to calculate the adjusted $R^2$ because it corrects for the tendency of $R^2$ to approach one as the number of explanatory variables increases (even though this effect is of little practical importance in large samples). The adjusted $R^2$ does this by assessing a degrees-of-freedom penalty for each variable added to the equation. It is always lower than the $R^2$ except in the case where only one variable is included in the model; then the two measures are equal. In Table 3.2 the adjusted $R^2$ is 0.041—only slightly smaller than the unadjusted $R^2$ of 0.042. The adjusted $R^2$ indicates that the model explains 4.3 percent of the variance in individual labor supply.

Finally, the formal measure of goodness of fit is given by the $F$-statistic. The $F$-statistic tests the null hypothesis that all of the slope coefficients are jointly equal to zero. Unless the model is very poor (has no statistically significant variables), the $F$-statistic is nearly always statistically significant, enabling the null hypothesis to be rejected. The model reported in Table 3.2 is no exception. The $F$-statistic of 42.16 greatly exceeds the tabulated value of $2.37$ (4 degrees of freedom in the numerator, 3818 degrees of freedom in the denominator, 95 percent confidence level).

**SPECIALIZED F-TESTS**

There are a variety of statistical tests that a researcher may wish to carry out besides assessing the significance of individual coefficients or the entire regression equation. For example, a researcher may be interested in determining the collective contribution of a set of variables to an equation or in testing whether two regression equations are significantly different. Both of these statistical tests can be carried out using an $F$-statistic.

**Testing the Contribution of a Subset of Variables**

Suppose that a researcher estimated a model of a person's income as a function of labor force variables, demographic variables, and health variables. For reasons that will become apparent in a moment, call this model the *unrestricted* model. Now suppose that the researcher would like to test whether the health variables collectively contribute to the regression equation. To conduct this test, the health variables are dropped from the model and the model is reestimated including only the labor force variables and the demographic variables. In essence, because the health variables have been dropped from the equation, their coefficients have been set to zero. Because of this, the second regression is called the *restricted* model.

It is always the case that dropping variables will cause the $R^2$ to fall and the sum of squared errors (SSE) to increase. Consequently, it is possible to test whether the health variables contribute to the regression by testing whether the difference in the
Regression Analysis

$R^2$s for the unrestricted and restricted models is statistically significant. If the unrestricted model contains $k$ variables and we drop $g$ variables, the appropriate $F$-statistic is as follows:

$$F = \frac{(R^2_{uk} - R^2_{uk-g})g}{(1 - R^2_{uk})/[(N - (k + 1)]}$$

This $F$-statistic has degrees of freedom equal to $g$ and $N - (k + 1)$. If the calculated value of the $F$-statistic is greater than the tabulated value, the null hypothesis of no difference between the restricted and unrestricted models is rejected—that is, the subset of variables does contribute to the regression equation.

Table 3.3 reports two models of total personal income. The unrestricted model includes age, education level, work status, two health-status variables (poor health status and the number of prior hospital stays in the previous year), and dummy indicator variables for marital status, gender, and race. All of these variables except hospital stay are statistically significant at the 95 percent confidence level or better. The restricted model contains the same variables, except that the health status variables are dropped. Using the $R^2$s for the two models, it is possible to test if the health variables explain a statistically significant portion of the variation in total personal income.

The unrestricted $R^2$ reported in Table 3.3 is 42.3 percent and the restricted $R^2$ is 42.1 percent. The Unrestricted $R^2$ is the $R^2$ associated with the original model (with all of the variables in Table 3.3). Because two variables were dropped, $g$ is equal to 2. $N$ is the sample size of 5,000 and $k$ is the number of explanatory variables (8) in the full (unrestricted) model. Thus, the calculated $F$-statistic is given by:

$$F = \frac{(0.4228 - 0.4208)^2}{(1 - 0.4228)/4991} = 8.647$$

The tabulated value for the $F$-statistic with df = 2,4991 at a 95 percent confidence level is 3.00. Since 8.647 is greater than 3.00, the null hypothesis that there is no difference between the unrestricted and restricted models is rejected. In other words, the health variables make a statistically significant contribution to the original model in Table 3.3. Of course, this should come as no surprise in the present example because the poor-health variable was statistically significant in the unrestricted model. Consequently, removing the poor-health variable, along with the prior-hospitalization variable, would be expected to result in a statistically significant drop in the model’s $R^2$. In some instances, however, the collective contribution of a group of variables might not be so obvious. For example, one could imagine having several health variables in the model, each of which was significant at around the 90 percent confidence level.
### Table 3.3
Income Level Regression with, and without, Health-Status Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Unrestricted Model</th>
<th></th>
<th>Restricted Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>t</td>
<td>Coefficient</td>
<td>t</td>
</tr>
<tr>
<td>Age</td>
<td>10.72</td>
<td>13.72</td>
<td>9.39</td>
<td>13.23</td>
</tr>
<tr>
<td>White</td>
<td>111.68</td>
<td>3.09</td>
<td>121.56</td>
<td>3.37</td>
</tr>
<tr>
<td>Education</td>
<td>17.99</td>
<td>8.80</td>
<td>19.26</td>
<td>9.51</td>
</tr>
<tr>
<td>Female</td>
<td>-472.36</td>
<td>-16.66</td>
<td>-472.02</td>
<td>-16.66</td>
</tr>
<tr>
<td>Married</td>
<td>137.41</td>
<td>4.59</td>
<td>138.96</td>
<td>4.64</td>
</tr>
<tr>
<td>Poor Health</td>
<td>-209.51</td>
<td>-4.18</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Hospital Stay</td>
<td>39.63</td>
<td>0.76</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Work</td>
<td>1147.90</td>
<td>34.66</td>
<td>1157.48</td>
<td>35.01</td>
</tr>
<tr>
<td>Constant</td>
<td>-124.56</td>
<td>-3.30</td>
<td>-129.79</td>
<td>-3.44</td>
</tr>
</tbody>
</table>

| R²            | 0.4228             |       | 0.4208           |       |
| N             | 5,000              |       | 5,000            |       |

*Source: Calculations based on the 1984 Panel (Waves 3&4) of the Survey of Income and Program Participation (SIPP).*

where none of the variables in question is statistically significant individually, for these variables to be statistically significant collectively.

### Testing the Equality of Regression Equations

Another important use of the $F$-statistic is to test the equality of regression equations based on different subsamples. For example, we might be interested in testing whether the determinants of personal income differ by gender. One way to test such a hypothesis is to construct dummy variables (variables which have values of only zero and one) to capture the effects of gender. Dummy variables were used in the models reported in Table 3.3; the interpretation of such variables is discussed in Chapter 4. Another method, however, is to split the sample into the groups defined by the dummy variable (i.e., gender), run separate regressions for each of the subsamples, and test the null hypothesis that the two regressions are equivalent.

To illustrate this approach, consider a simple model hypothesizing that income levels are a function of marital status, age, education level, race, poor health, and work status. The first step is to estimate the model of income level using all of the
observation; for both men and women. From this model, obtain the sum of squared errors (SSE).

Next the sample is split by gender, the same model is estimated (with the same variables as the full model) for both men and women, and the sum of squared errors SSE_m and SSE_w is obtained for each model. If the sample of women contained N observations, that for men contained M observations, and each of the equations contained k explanatory variables, the relevant F-test would be:

\[ F_{(k-1, N-M-2(k-1))} = \frac{\hat{SSE}_u - (\hat{SSE}_m + \hat{SSE}_w) / (k+1)}{\hat{SSE}_u / (N+M-2(k+1))} \]

This F-statistic is distributed with k + 1 and N + M - 2 (k + 1) degrees of freedom. If the calculated value of the F-statistic is greater than the tabulated value for a given level of statistical significance, the null hypothesis that the determinants of hours worked are the same for men and women is rejected.4

Table 3.4 reports the regression results to carry out this F-test using a sample of 5,000 observations from the 1984 Panel of the Survey of Income and Program Participation. The first two columns of Table 3.4 contain the estimated coefficients and t-statistics for the pooled sample of men and women. Columns 3 and 4 present the model for men and columns 5 and 6 present the model for women. From visual inspection of the coefficient estimates and t-statistics, it appears that the income models for men and women are different. This can be tested using the SSE values reported at the bottom of the table for each model. The SSE for the pooled model is 4,985 x 10^6; the SSE_m is 7,527 x 10^4; and the SSE_w is 3,626 x 10^6. The number of explanatory variables, k, is 6.

Substituting these values into the above formula for the F-test yields:

\[ F_{3,4986} = \frac{[4,985 x 10^6 - (7,527 x 10^4 + 3,626 x 10^6)]/7}{(7,527 x 10^4 + 3,626 x 10^6)/4986} = 98.63 \]

The critical value for the F-statistic with df = 7,4986 is 2.01 at the 95% confidence level. Therefore, the calculated value of F exceeds the tabulated value and the null hypothesis that there is no difference in the regression equations of income levels for men and women is rejected. In short, the determinants of male and female income levels, as indicated by the regression equations, are different.

ASSUMPTIONS UNDERLYING REGRESSION MODELS

It is very important to recognize that regression models are based on a number of assumptions that have implications for the interpretation and validity of the
### Table 3.4

**Income Level Regressions for Male, Female, and Pooled Samples**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Pooled Sample</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>9.46</td>
<td>11.85</td>
<td>7.69</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>12.98</td>
<td>11.06</td>
</tr>
<tr>
<td>White</td>
<td>80.37</td>
<td>2.17</td>
<td>7.91</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>0.26</td>
<td>214.50</td>
</tr>
<tr>
<td>Education</td>
<td>15.15</td>
<td>7.25</td>
<td>14.57</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>8.81</td>
<td>22.58</td>
</tr>
<tr>
<td>Married</td>
<td>191.28</td>
<td>6.27</td>
<td>-159.92</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>-6.55</td>
<td>498.41</td>
</tr>
<tr>
<td>Poor Health</td>
<td>-207.86</td>
<td>-4.15</td>
<td>-84.49</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>-2.32</td>
<td>-426.83</td>
</tr>
<tr>
<td>Work</td>
<td>1228.60</td>
<td>36.53</td>
<td>854.79</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>32.64</td>
<td>1325.90</td>
</tr>
<tr>
<td>Constant</td>
<td>-317.74</td>
<td>-8.62</td>
<td>-142.23</td>
</tr>
<tr>
<td></td>
<td>t</td>
<td>-4.34</td>
<td>-514.52</td>
</tr>
</tbody>
</table>

| $R^2$ | 0.39 | 0.45 | 0.44 |
| F    | 533.64 | 343.22 | 325.89 |
| SSE  | $4.985 \times 10^8$ | $7.527 \times 10^8$ | $3.626 \times 10^9$ |
| N    | 5,000 | 2,476 | 2,524 |

Source: Calculations based on the 1984 Panel of the Survey of Income and Program Participation (SIPP).

Results obtained. When these assumptions are met, the estimates produced by ordinary least squares (OLS) regression have a number of desirable statistical properties. On the other hand, when the assumptions are not met, the OLS estimators no longer have these desirable statistical properties. Before we proceed further, therefore, these assumptions and their implications need to be identified.

### No Specification Error

Actually, this is not one assumption but several. Unfortunately, there are many possible categories of specification error. Regression analysis assumes that the relationship between the dependent variable and the explanatory variables is linear in the coefficients. At first glance, this appears to be an extremely restrictive assumption. However, in Chapter 4 it will be shown that, with suitable transformations to the data, regression analysis can be used to estimate a broad range of nonlinear relationships. Failure to take account of nonlinear relationships between the explanatory and dependent variables results in biased coefficient estimates.

Another type of specification error is the omission of relevant explanatory variables. Like the failure to estimate the proper functional form, excluded variables result in biased coefficient estimates. In truth, this assumption is nearly
always violated to some degree. After all, the basic rationale for building a model in the first place is to simplify reality so that it can be better understood. The practical issue then, is whether all of the important variables have been included in the model so as to minimize the bias introduced by excluded variables.

Relevant variables are excluded from analysis for a variety of reasons. Sometimes a researcher knows that a particular variable should be included in the model (i.e., the variable is part of the theoretical model) but data on the variable are unavailable. In this instance, the researcher is at least aware that the variable is missing and that some bias has probably been introduced into the coefficients that have been estimated. Usually, in such a case, the researcher will try to minimize the bias by including a variable in the model that is a proxy for the one that is missing. A more serious problem is the omission of a variable because the researcher does not realize that it belongs in the theoretical model. In such a case, the researcher may unknowingly place confidence in an estimated model that has misleading policy implications because the coefficients are biased.

The inclusion of irrelevant variables is less serious than the problem of excluded variables. In models containing irrelevant variables, it can be shown that the coefficient estimates are unbiased but not efficient. The estimated standard errors are also unbiased. Consequently, t-tests of the statistical significance of the coefficients remain valid in such models.

No Measurement Error

Measurement errors are assumed to be small in magnitude and random. One would expect systematic measurement errors to result in biased estimates of regression coefficients and standard errors. However, even random measurement errors can lead to biased estimates when such errors occur in measuring the explanatory variables. It is extremely difficult to decipher the direction and magnitude of the bias introduced by measurement error in multivariate regression models. Much of the work in modern survey design is focused on reducing measurement error at its source by improving the quality of data collected.

No Perfect Multicollinearity

It is also assumed that there is no perfect correlation among the explanatory variables. High correlations among two or more of the explanatory variables create a problem known as multicollinearity. Multicollinearity makes it impossible for the regression model to decompose the variation of the dependent variable that is due to one variable versus another. As a consequence, one gets highly unstable coefficients and t-statistics that fluctuate widely in magnitude (and even sign) depending upon which variables are included in the equation.
Assumptions about the Error Term

Most of the assumptions underlying the use of the regression model concern the distribution of the error term. As discussed earlier in reference to Figure 3.1, the estimated regression line will not be a perfect predictor of the actual values of the dependent variable. Sometimes the regression will overpredict the actual value; other times it will underpredict. There are several possible reasons for these errors—improper functional form, measurement errors, and excluded variables. To this list we might add the unpredictability of human behavior.

The regression residuals are assumed to be normally distributed with mean zero and constant variance. The assumption of normality is not needed to derive the formulas for the regression coefficients and standard errors. Consequently, even if the residuals are not normally distributed, the regression coefficients are still the best linear unbiased estimators (BLUE). On the other hand, the t-tests for the significance of the coefficients do depend on the assumption of normality. Fortunately, it can be shown that the sampling distribution for the least squares estimators approaches the normal distribution for large sample sizes. Even in small samples the least squares estimates may not be too seriously affected if the distribution of the residuals is not too different from the normal distribution.

A second assumption concerning the residuals is that they have an expected value of zero. Violating the assumption that the expected value of the residuals is equal to zero affects only the constant term—shifting it by the (nonzero) mean of the errors. This is often not a major concern because the slope coefficients are usually the main focus of interest in policy models. In fact, one of the virtues of the constant term is that it tends to absorb the effects of various violations of assumptions, thereby helping to reduce the bias that would otherwise be introduced into the slope coefficients.

Third, the error terms are assumed to have a constant variance or to be homoscedastic. For a given set of values for the explanatory variables, a regression model will predict one value for the dependent variable. However, corresponding to this prediction, there is a distribution of actual values for the dependent variable and, therefore, a distribution of errors. The assumption of homoscedasticity says that the variance of these errors for different values of the explanatory variables should be constant. If the variance of the residuals is not constant, the errors are said to be heteroscedastic. In the presence of heteroscedasticity the regression coefficients are still unbiased, but the t-statistics are unreliable.

Fourth, it is assumed that there is no correlation among the residuals. The presence of such an association, known as autocorrelation, is primarily a problem that occurs with time-series data. In cross-sectional data, however, the presence of autocorrelation can indicate excluded variables or other specification problems. As in the case of heteroscedasticity, the regression coefficients are unbiased in the presence of autocorrelation, but the t-statistics are unreliable.
Fifth, the residuals are assumed to be uncorrelated with the independent variables. Correlation of the error term with the explanatory variables introduces bias in the coefficient estimates.

The overarching assumption of the regression model is that the errors are random. It is assumed, for example, that the number of excluded variables is numerous so that the individual effects of any particular omitted variable are negligible. Similarly, measurement errors are assumed to be negligible and unpredictable. And, of course, human behavior is assumed to have a random component that cannot be systematically described by a regression model. To the extent that the assumption of randomness is not met, it implies that other violations of assumptions may be present. The statistical implications will depend upon which of the other assumptions is violated. For example, a systematic pattern in the errors that was due to the choice of an improper functional form or omitted variables will result in biased coefficient estimates. On the other hand, autocorrelation of residuals due to the use of time-series data suggests that the $t$-statistics associated with the variables cannot be trusted.

Subsequent chapters will examine how to test for violations of these assumptions. Given the long list of assumptions just outlined, it would seem impossible to conduct a regression analysis without violating at least one of the assumptions—and this is the case. Yet in many instances it is possible to correct these violations and still continue to use regression analysis. When such corrections are not possible, however, more advanced statistical techniques must be used.

There is disagreement in the literature about how serious violations of the regression assumptions actually are. This is not a meaningful discussion in the abstract. In any particular application, violation of one or more of the assumptions may introduce substantial bias or seriously undermine the validity of statistical tests. The same types of violations in another application might have comparatively mild repercussions. There has been considerable recent work on robust estimators, which are less sensitive to violations of assumptions.  

NOTES

1. Further details on the procedure used to solve for the regression coefficient formulas (as well as the other results presented in this chapter) can be found in any econometrics text and most specialized texts on regression. See, for example, Maddala (1977), Pindyck and Rubinfeld (1991), Ghosh (1991), Hamilton (1992), and Greene (1993).

2. For an excellent reference of the estimation of labor supply equations, see Killingsworth (1983).

3. There are equivalent versions of this $F$-test based on the regression sum of squares (RSS) and the sum of squared errors (SSE). For example, see Johnson, Johnson, and Buse (1987). Godfrey (1988) notes that this test amounts to a test for omitted variables if one assumes that the full model is correct.

4. This test was suggested by Chow (1960).

5. The reader will note that the form of the $F$-test in each of the above applications is very similar. There are, in fact, many such tests. A good introductory discussion is given in Johnson, Johnson, and Buse (1987).