5.9 REGRESSION ANALYSIS AND ANALYSIS OF VARIANCE

In this section we study regression analysis from the point of view of the analysis of variance and introduce the reader to an illuminating and complementary way of looking at the statistical inference problem.

In Chapter 3, Section 3.5, we developed the following identity:

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 = \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2 \qquad \textbf{(3.5.2)}$$

that is, TSS = ESS + RSS, which decomposed the total sum of squares (TSS) into two components: explained sum of squares (ESS) and residual sum of squares (RSS). A study of these components of TSS is known as the **analysis of variance** (ANOVA) from the regression viewpoint.

Associated with any sum of squares is its df, the number of independent observations on which it is based. TSS has $n - 1$ df because we lose 1 df in computing the sample mean $\bar{Y}$. RSS has $n - 2$ df. (Why?) (*Note:* This is true only for the two-variable regression model with the intercept β_1 present.) ESS has 1 df (again true of the two-variable case only), which follows from the fact that ESS $= \hat{\beta}_2^2 \sum x_i^2$ is a function of $\hat{\beta}_2$ only, since $\sum x_i^2$ is known.

Let us arrange the various sums of squares and their associated df in Table 5.3, which is the standard form of the AOV table, sometimes called the **ANOVA table.** Given the entries of Table 5.3, we now consider the following variable:

$$\begin{aligned} F &= \frac{\text{MSS of ESS}}{\text{MSS of RSS}} \\ &= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum \hat{u}_i^2/(n-2)} \qquad \textbf{(5.9.1)} \\ &= \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2} \end{aligned}$$

If we assume that the disturbances u_i are normally distributed, which we do under the CNLRM, and if the null hypothesis (H_0) is that $\beta_2 = 0$, then it can be shown that the F variable of (5.9.1) follows the F distribution with

TABLE 5.3 ANOVA TABLE FOR THE TWO-VARIABLE REGRESSION MODEL

Source of variation	SS*	df	MSS†
Due to regression (ESS)	$\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	1	$\hat{\beta}_2^2 \sum x_i^2$
Due to residuals (RSS)	$\sum \hat{u}_i^2$	$n - 2$	$\frac{\sum \hat{u}_i^2}{n-2} = \hat{\sigma}^2$
TSS	$\sum y_i^2$	$n - 1$	

*SS means sum of squares.
†Mean sum of squares, which is obtained by dividing SS by their df.

1 df in the numerator and $(n - 2)$ df in the denominator. (See Appendix 5A, Section 5A.3, for the proof. The general properties of the F distribution are discussed in **Appendix A**.)

What use can be made of the preceding F ratio? It can be shown[18] that

$$E\left(\hat{\beta}_2^2 \sum x_i^2\right) = \sigma^2 + \beta_2^2 \sum x_i^2 \qquad \textbf{(5.9.2)}$$

and

$$E\frac{\sum \hat{u}_i^2}{n-2} = E(\hat{\sigma}^2) = \sigma^2 \qquad \textbf{(5.9.3)}$$

(Note that β_2 and σ^2 appearing on the right sides of these equations are the true parameters.) Therefore, if β_2 is in fact zero, Eqs. (5.9.2) and (5.9.3) both provide us with identical estimates of true σ^2. In this situation, the explanatory variable X has no linear influence on Y whatsoever and the entire variation in Y is explained by the random disturbances u_i. If, on the other hand, β_2 is not zero, (5.9.2) and (5.9.3) will be different and part of the variation in Y will be ascribable to X. Therefore, the F ratio of (5.9.1) provides a test of the null hypothesis H_0: $\beta_2 = 0$. Since all the quantities entering into this equation can be obtained from the available sample, this F ratio provides a test statistic to test the null hypothesis that true β_2 is zero. All that needs to be done is to compute the F ratio and compare it with the critical F value obtained from the F tables at the chosen level of significance, or obtain the ***p* value** of the computed F statistic.

To illustrate, let us continue with our consumption–income example. The ANOVA table for this example is as shown in Table 5.4. The computed F value is seen to be 202.87. The p value of this F statistic corresponding to 1 and 8 df cannot be obtained from the F table given in **Appendix D**, but by using electronic statistical tables it can be shown that the p value is 0.0000001, an extremely small probability indeed. If you decide to choose the level-of-significance approach to hypothesis testing and fix α at 0.01, or a 1 percent level, you can see that the computed F of 202.87 is obviously significant at this level. Therefore, if we reject the null hypothesis that $\beta_2 = 0$, the probability of committing a Type I error is very small. For all practical

TABLE 5.4 ANOVA TABLE FOR THE CONSUMPTION–INCOME EXAMPLE

Source of variation	SS	df	MSS	
Due to regression (ESS)	8552.73	1	8552.73	$F = \dfrac{8552.73}{42.159}$
Due to residuals (RSS)	337.27	8	42.159	$= 202.87$
TSS	8890.00	9		

[18] For proof, see K. A. Brownlee, *Statistical Theory and Methodology in Science and Engineering*, John Wiley & Sons, New York, 1960, pp. 278–280.

purposes, our sample could not have come from a population with zero β_2 value and we can conclude with great confidence that X, income, does affect Y, consumption expenditure.

Refer to Theorem 5.7 of Appendix 5A.1, which states that the square of the t value with k df is an F value with 1 df in the numerator and k df in the denominator. For our consumption–income example, if we assume H_0: $\beta_2 = 0$, then from (5.3.2) it can be easily verified that the estimated t value is 14.26. This t value has 8 df. Under the same null hypothesis, the F value was 202.87 with 1 and 8 df. Hence $(14.24)^2 = F$ value, except for the rounding errors.

Thus, the t and the F tests provide us with two alternative but complementary ways of testing the null hypothesis that $\beta_2 = 0$. If this is the case, why not just rely on the t test and not worry about the F test and the accompanying analysis of variance? For the two-variable model there really is no need to resort to the F test. But when we consider the topic of multiple regression we will see that the F test has several interesting applications that make it a very useful and powerful method of testing statistical hypotheses.

5.10 APPLICATION OF REGRESSION ANALYSIS: THE PROBLEM OF PREDICTION

On the basis of the sample data of Table 3.2 we obtained the following sample regression:

$$\hat{Y}_i = 24.4545 + 0.5091X_i \tag{3.6.2}$$

where $\hat{Y}_i$ is the estimator of true $E(Y_i)$ corresponding to given X. What use can be made of this **historical regression?** One use is to "predict" or "forecast" the future consumption expenditure Y corresponding to some given level of income X. Now there are two kinds of predictions: (1) prediction of the conditional mean value of Y corresponding to a chosen X, say, X_0, that is the point on the population regression line itself (see Figure 2.2), and (2) prediction of an individual Y value corresponding to X_0. We shall call these two predictions the **mean prediction** and **individual prediction.**

Mean Prediction[19]

To fix the ideas, assume that $X_0 = 100$ and we want to predict $E(Y \mid X_0 = 100)$. Now it can be shown that the historical regression (3.6.2) provides the point estimate of this mean prediction as follows:

$$\begin{aligned}\hat{Y}_0 &= \hat{\beta}_1 + \hat{\beta}_2 X_0 \\ &= 24.4545 + 0.5091(100) \qquad \textbf{(5.10.1)} \\ &= 75.3645\end{aligned}$$

[19] For the proofs of the various statements made, see App. 5A, Sec. 5A.4.

where $\hat{Y}_0 =$ estimator of $E(Y \mid X_0)$. It can be proved that this point predictor is a best linear unbiased estimator (BLUE).

Since $\hat{Y}_0$ is an estimator, it is likely to be different from its true value. The difference between the two values will give some idea about the prediction or forecast error. To assess this error, we need to find out the sampling distribution of $\hat{Y}_0$. It is shown in Appendix 5A, Section 5A.4, that $\hat{Y}_0$ in Eq. (5.10.1) is normally distributed with mean $(\beta_1 + \beta_2 X_0)$ and the variance is given by the following formula:

$$\operatorname{var}(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right] \tag{5.10.2}$$

By replacing the unknown σ^2 by its unbiased estimator $\hat{\sigma}^2$, we see that the variable

$$t = \frac{\hat{Y}_0 - (\beta_1 + \beta_2 X_0)}{\operatorname{se}(\hat{Y}_0)} \tag{5.10.3}$$

follows the t distribution with $n - 2$ df. The t distribution can therefore be used to derive confidence intervals for the true $E(Y_0 \mid X_0)$ and test hypotheses about it in the usual manner, namely,

$$\Pr[\hat{\beta}_1 + \hat{\beta}_2 X_0 - t_{\alpha/2}\operatorname{se}(\hat{Y}_0) \le \beta_1 + \beta_2 X_0 \le \hat{\beta}_1 + \hat{\beta}_2 X_0 + t_{\alpha/2}\operatorname{se}(\hat{Y}_0)] = 1 - \alpha \tag{5.10.4}$$

where $\operatorname{se}(\hat{Y}_0)$ is obtained from (5.10.2).

For our data (see Table 3.3),

$$\begin{aligned}\operatorname{var}(\hat{Y}_0) &= 42.159\left[\frac{1}{10} + \frac{(100 - 170)^2}{33{,}000}\right]\\ &= 10.4759\end{aligned}$$

and

$$\operatorname{se}(\hat{Y}_0) = 3.2366$$

Therefore, the 95% confidence interval for true $E(Y \mid X_0) = \beta_1 + \beta_2 X_0$ is given by

$$75.3645 - 2.306(3.2366) \le E(Y_0 \mid X = 100) \le 75.3645 + 2.306(3.2366)$$

that is,

$$67.9010 \le E(Y \mid X = 100) \le 82.8381 \tag{5.10.5}$$

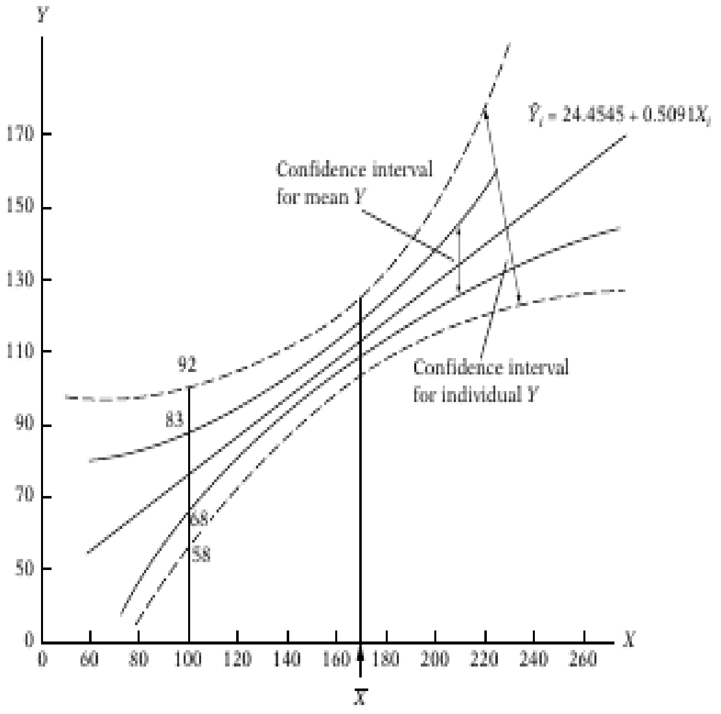

FIGURE 5.6 Confidence intervals (bands) for mean *Y* and individual *Y* values.

Thus, given $X_0 = 100$, in repeated sampling, 95 out of 100 intervals like (5.10.5) will include the true mean value; the single best estimate of the true mean value is of course the point estimate 75.3645.

If we obtain 95% confidence intervals like (5.10.5) for each of the *X* values given in Table 3.2, we obtain what is known as the **confidence interval,** or **confidence band,** for the population regression function, which is shown in Figure 5.6.

Individual Prediction

If our interest lies in predicting an individual *Y* value, Y_0, corresponding to a given *X* value, say, X_0, then, as shown in Appendix 5, Section 5A.3, a best linear unbiased estimator of Y_0 is also given by (5.10.1), but its variance is as follows:

$$\operatorname{var}(Y_0 - \hat{Y}_0) = E[Y_0 - \hat{Y}_0]^2 = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right] \qquad \textbf{(5.10.6)}$$

It can be shown further that Y_0 also follows the normal distribution with mean and variance given by (5.10.1) and (5.10.6), respectively. Substituting $\hat{\sigma}^2$

for the unknown σ^2, it follows that

$$t = \frac{Y_0 - \hat{Y}_0}{\text{se}\,(Y_0 - \hat{Y}_0)}$$

also follows the t distribution. Therefore, the t distribution can be used to draw inferences about the true Y_0. Continuing with our consumption–income example, we see that the point prediction of Y_0 is 75.3645, the same as that of $\hat{Y}_0$, and its variance is 52.6349 (the reader should verify this calculation). Therefore, the 95% confidence interval for Y_0 corresponding to $X_0 = 100$ is seen to be

$$(58.6345 \le Y_0 \mid X_0 = 100 \le 92.0945) \qquad \textbf{(5.10.7)}$$

Comparing this interval with (5.10.5), we see that the confidence interval for individual Y_0 is wider than that for the mean value of Y_0. (Why?) Computing confidence intervals like (5.10.7) conditional upon the X values given in Table 3.2, we obtain the 95% confidence band for the individual Y values corresponding to these X values. This confidence band along with the confidence band for $\hat{Y}_0$ associated with the same X's is shown in Figure 5.6.

Notice an important feature of the confidence bands shown in Figure 5.6. The width of these bands is smallest when $X_0 = \bar{X}$. (Why?) However, the width widens sharply as X_0 moves away from $\bar{X}$. (Why?) This change would suggest that the predictive ability of the *historical* sample regression line falls markedly as X_0 departs progressively from $\bar{X}$. **Therefore, one should exercise great caution in "extrapolating" the historical regression line to predict $E(Y \mid X_0)$ or Y_0 associated with a given X_0 that is far removed from the sample mean $\bar{X}$.**

5.11 REPORTING THE RESULTS OF REGRESSION ANALYSIS

There are various ways of reporting the results of regression analysis, but in this text we shall use the following format, employing the consumption–income example of Chapter 3 as an illustration:

$$\begin{aligned}
\hat{Y}_i &= 24.4545 \quad + \quad 0.5091X_i \\
\text{se} &= (6.4138) \qquad (0.0357) \qquad\qquad r^2 = 0.9621 \\
t &= (3.8128) \qquad (14.2605) \qquad\qquad \text{df} = 8 \\
p &= (0.002571) \qquad (0.000000289) \qquad F_{1,8} = 202.87
\end{aligned} \qquad \textbf{(5.11.1)}$$

In Eq. (5.11.1) the figures in the first set of parentheses are the estimated standard errors of the regression coefficients, the figures in the second set are estimated t values computed from (5.3.2) under the null hypothesis that the true population value of each regression coefficient individually is zero (e.g., $3.8128 = 24.4545 \div 6.4138$), and the figures in the third set are the estimated p values. Thus, for 8 df the probability of obtaining a t value of 3.8128 or greater is 0.0026 and the probability of obtaining a t value of 14.2605 or larger is about 0.0000003.

By presenting the p values of the estimated t coefficients, we can see at once the exact level of significance of each estimated t value. Thus, under the null hypothesis that the true population intercept value is zero, the exact probability (i.e., the p value) of obtaining a t value of 3.8128 or greater is only about 0.0026. Therefore, if we reject this null hypothesis, the probability of our committing a Type I error is about 26 in 10,000, a very small probability indeed. For all practical purposes we can say that the true population intercept is different from zero. Likewise, the p value of the estimated slope coefficient is zero for all practical purposes. If the true MPC were in fact zero, our chances of obtaining an MPC of 0.5091 would be practically zero. Hence we can reject the null hypothesis that the true MPC is zero.

Earlier we showed the intimate connection between the F and t statistics, namely, $F_{1,k} = t_k^2$. Under the null hypothesis that the true $\beta_2 = 0$, (5.11.1) shows that the F value is 202.87 (for 1 numerator and 8 denominator df) and the t value is about 14.24 (8 df); as expected, the former value is the square of the latter value, except for the roundoff errors. The ANOVA table for this problem has already been discussed.

5.12 EVALUATING THE RESULTS OF REGRESSION ANALYSIS

In Figure I.4 of the Introduction we sketched the anatomy of econometric modeling. Now that we have presented the results of regression analysis of our consumption–income example in (5.11.1), we would like to question the adequacy of the fitted model. How “good” is the fitted model? We need some criteria with which to answer this question.

First, are the signs of the estimated coefficients in accordance with theoretical or prior expectations? A priori, β_2, the marginal propensity to consume (MPC) in the consumption function, should be positive. In the present example it is. Second, if theory says that the relationship should be not only positive but also statistically significant, is this the case in the present application? As we discussed in Section 5.11, the MPC is not only positive but also statistically significantly different from zero; the p value of the estimated t value is extremely small. The same comments apply about the intercept coefficient. Third, how well does the regression model explain variation in the consumption expenditure? One can use r^2 to answer this question. In the present example r^2 is about 0.96, which is a very high value considering that r^2 can be at most 1.

Thus, the model we have chosen for explaining consumption expenditure behavior seems quite good. But before we sign off, we would like to find out whether our model satisfies the assumptions of CNLRM. We will not look at the various assumptions now because the model is patently so simple. But there is one assumption that we would like to check, namely, the normality of the disturbance term, u_i. Recall that the *t* and *F* tests used before require that the error term follow the normal distribution. Otherwise, the testing procedure will not be valid in small, or finite, samples.

Normality Tests

Although several tests of normality are discussed in the literature, we will consider just three: (1) histogram of residuals; (2) normal probability plot (NPP), a graphical device; and (3) the **Jarque–Bera** test.

Histogram of Residuals. A histogram of residuals is a simple graphic device that is used to learn something about the shape of the PDF of a random variable. On the horizontal axis, we divide the values of the variable of interest (e.g., OLS residuals) into suitable intervals, and in each class interval we erect rectangles equal in height to the number of observations (i.e., frequency) in that class interval. If you mentally superimpose the bell-shaped normal distribution curve on the histogram, you will get some idea as to whether normal (PDF) approximation may be appropriate. A concrete example is given in Section 5.13 (see Figure 5.8). It is always a good practice to plot the histogram of the residuals as a rough and ready method of testing for the normality assumption.

Normal Probability Plot. A comparatively simple graphical device to study the shape of the probability density function (PDF) of a random variable is the **normal probability plot (NPP)** which makes use of *normal probability paper*, a specially designed graph paper. On the horizontal, or *x*, axis, we plot values of the variable of interest (say, OLS residuals, $\hat{u}_i$), and on the vertical, or *y*, axis, we show the expected value of this variable if it were normally distributed. Therefore, if the variable is in fact from the normal population, the NPP will be approximately a straight line. The NPP of the residuals from our consumption–income regression is shown in Figure 5.7, which is obtained from the MINITAB software package, version 13. As noted earlier, if the fitted line in the NPP is approximately a straight line, one can conclude that the variable of interest is normally distributed. In Figure 5.7, we see that residuals from our illustrative example are approximately normally distributed, because a straight line seems to fit the data reasonably well.

MINITAB also produces the **Anderson–Darling normality test,** known as the $\boldsymbol{A^2}$ **statistic.** The underlying null hypothesis is that the variable under consideration is normally distributed. As Figure 5.7 shows, for our example, the computed A^2 statistic is 0.394. The *p value* of obtaining such a value of A^2 is 0.305, which is reasonably high. Therefore, we do not reject the

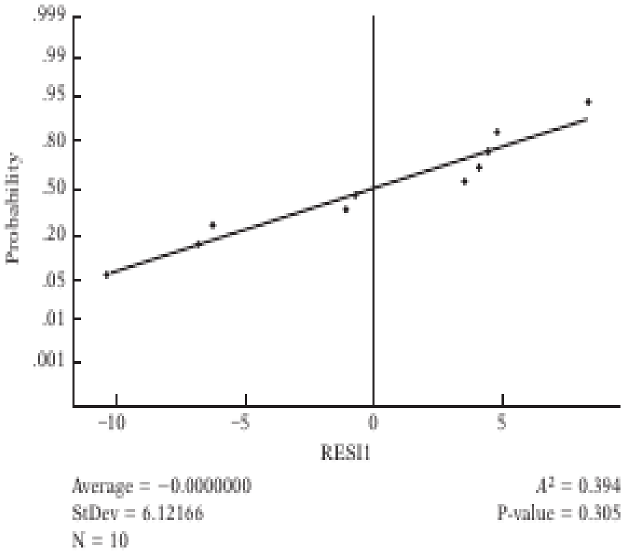

FIGURE 5.7 Residuals from consumption–income regression.

hypothesis that the residuals from our consumption–income example are normally distributed. Incidentally, Figure 5.7 shows the parameters of the (normal) distribution, the mean is approximately 0 and the standard deviation is about 6.12.

Jarque–Bera (JB) Test of Normality.[20] The JB test of normality is an *asymptotic,* or large-sample, test. It is also based on the OLS residuals. This test first computes the **skewness** and **kurtosis** (discussed in **Appendix A**) measures of the OLS residuals and uses the following test statistic:

$$\text{JB} = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right] \qquad \textbf{(5.12.1)}$$

where n = sample size, S = skewness coefficient, and K = kurtosis coefficient. For a normally distributed variable, $S = 0$ and $K = 3$. Therefore, the JB test of normality is a test of the joint hypothesis that S and K are 0 and 3, respectively. In that case the value of the JB statistic is expected to be 0.

Under the null hypothesis that the residuals are normally distributed, Jarque and Bera showed that *asymptotically (i.e., in large samples) the JB statistic given in (5.12.1) follows the chi-square distribution with 2 df.* If the computed p value of the JB statistic in an application is sufficiently low, which will happen if the value of the statistic is very different from 0, one can reject the hypothesis that the residuals are normally distributed. But if

[20] See C. M. Jarque and A. K. Bera, "A Test for Normality of Observations and Regression Residuals," *International Statistical Review*, vol. 55, 1987, pp. 163–172.

the *p* value is reasonably high, which will happen if the value of the statistic is close to zero, we do not reject the normality assumption.

The sample size in our consumption–income example is rather small. Hence, strictly speaking one should not use the JB statistic. If we mechanically apply the JB formula to our example, the JB statistic turns out to be 0.7769. The *p* value of obtaining such a value from the chi-square distribution with 2 df is about 0.68, which is quite high. In other words, we may not reject the normality assumption for our example. Of course, bear in mind the warning about the sample size.

Other Tests of Model Adequacy

Remember that the CNLRM makes many more assumptions than the normality of the error term. As we examine econometric theory further, we will consider several tests of model adequacy (see Chapter 13). Until then, keep in mind that our regression modeling is based on several simplifying assumptions that may not hold in each and every case.

A CONCLUDING EXAMPLE

Let us return to Example 3.2 about food expenditure in India. Using the data given in (3.7.2) and adopting the format of (5.11.1), we obtain the following expenditure equation:

$$\begin{aligned}
\widehat{\text{FoodExp}}_i &= 94.2087 \quad + \quad 0.4368\ \text{TotalExp}_i \\
\text{se} &= (50.8563) \qquad (0.0783) \\
t &= (1.8524) \qquad (5.5770) \\
p &= (0.0695) \qquad (0.0000)^* \\
r^2 &= 0.3698; \qquad \text{df} = 53 \\
F_{1,53} &= 31.1034 \qquad (p\ \text{value} = 0.0000)^*
\end{aligned} \tag{5.12.2}$$

where * denotes extremely small.

First, let us interpret this regression. As expected, there is a positive relationship between expenditure on food and total expenditure. If total expenditure went up by a rupee, on average, expenditure on food increased by about 44 paise. If total expenditure were zero, the average expenditure on food would be about 94 rupees. Of course, this mechanical interpretation of the intercept may not make much economic sense. The r^2 value of about 0.37 means that 37 percent of the variation in food expenditure is explained by total expenditure, a proxy for income.

Suppose we want to test the null hypothesis that there is no relationship between food expenditure and total expenditure, that is, the true slope coefficient $\beta_2 = 0$. The estimated value of β_2 is 0.4368. If the null hypothesis were true, what is the probability of obtaining a value of 0.4368? Under the null hypothesis, we observe from (5.12.2) that the *t* value is 5.5770 and the *p* value of obtaining such a *t* value is practically zero. In other words, we can reject the null hypothesis resoundingly. But suppose the null hypothesis were that $\beta_2 = 0.5$. Now what? Using the *t* test we obtain:

$$t = \frac{0.4368 - 0.5}{0.0783} = -0.8071$$

The probability of obtaining a $|t|$ of 0.8071 is greater than 20 percent. Hence we do not reject the hypothesis that the true β_2 is 0.5.

Notice that, under the null hypothesis, the true slope coefficient is zero, the *F* value is 31.1034, as shown in (5.12.2). Under the same null hypothesis, we obtained a *t* value of 5.5770. If we square this value, we obtain 31.1029, which is about the same as the *F* value, again showing the close relationship between the *t* and the *F* statistic. (*Note:* The numerator df for the *F* statistic must be 1, which is the case here.)

Using the estimated residuals from the regression, what can we say about the probability distribution of the error term? The information is given in Figure 5.8. As the figure shows, the residuals from the food expenditure regression seem to be symmetrically distributed. Application of the Jarque–Bera test shows that the JB statistic is about 0.2576, and the probability of obtaining such a statistic under the normality assumption is about

(*Continued*)

A CONCLUDING EXAMPLE (*Continued*)

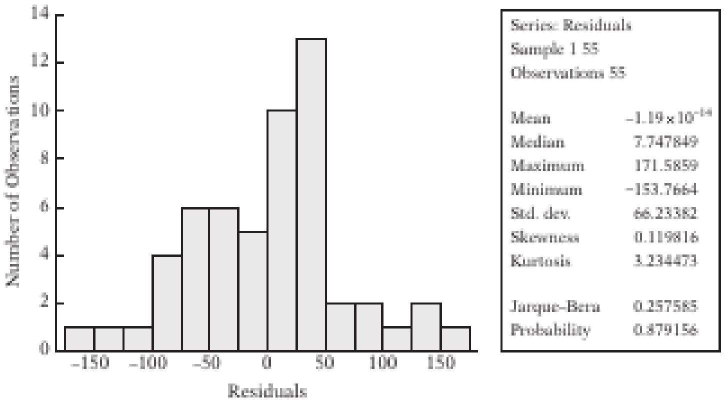

FIGURE 5.8 Residuals from the food expenditure regression.

88 percent. Therefore, we do not reject the hypothesis that the error terms are normally distributed. But keep in mind that the sample size of 55 observations may not be large enough.

We leave it to the reader to establish confidence intervals for the two regression coefficients as well as to obtain the normal probability plot and do mean and individual predictions.