

equation (b_1). The data contained in the DATAfile named *Beta* provides the total return (capital appreciation plus dividends) over 36 months for eight widely traded common stocks and the S&P 500.

The value of beta for the stock market will always be 1; thus, stocks that tend to rise and fall with the stock market will also have a beta close to 1. Betas greater than 1 indicate that the stock is more volatile than the market, and betas less than 1 indicate that the stock is less volatile than the market. For instance, if a stock has a beta of 1.4, it is 40% *more* volatile than the market, and if a stock has a beta of .4, it is 60% *less* volatile than the market.

Managerial Report

You have been assigned to analyze the risk characteristics of these stocks. Prepare a report that includes but is not limited to the following items.

- Compute descriptive statistics for each stock and the S&P 500. Comment on your results. Which stocks are the most volatile?
- Compute the value of beta for each stock. Which of these stocks would you expect to perform best in an up market? Which would you expect to hold their value best in a down market?
- Comment on how much of the return for the individual stocks is explained by the market.

Problem 2 U.S. Department of Transportation

As part of a study on transportation safety, the U.S. Department of Transportation collected data on the number of fatal accidents per 1000 licenses and the percentage of licensed drivers under the age of 21 in a sample of 42 cities. Data collected over a one-year period follow. These data are contained in the DATAfile named *Safety*.

Percent Under 21	Fatal Accidents per 1000 Licenses	Percent Under 21	Fatal Accidents per 1000 Licenses
13	2.962	17	4.100
12	0.708	8	2.190
8	0.885	16	3.623
12	1.652	15	2.623
11	2.091	9	0.835
17	2.627	8	0.820
18	3.830	14	2.890
8	0.368	8	1.267
13	1.142	15	3.224
8	0.645	10	1.014
9	1.028	10	0.493
16	2.801	14	1.443
12	1.405	18	3.614
9	1.433	10	1.926
10	0.039	14	1.643
9	0.338	16	2.943
11	1.849	12	1.913
12	2.246	15	2.814
14	2.855	13	2.634
14	2.352	9	0.926
14	1.294	17	3.256

Managerial Report

1. Develop numerical and graphical summaries of the data.
2. Use regression analysis to investigate the relationship between the number of fatal accidents and the percentage of drivers under the age of 21. Discuss your findings.
3. What conclusion and recommendations can you derive from your analysis?

Selecting a Point-and-Shoot Digital Camera

Consumer Reports tested 166 different point-and-shoot digital cameras. Based upon factors such as the number of megapixels, weight (oz.), image quality, and ease of use, they developed an overall score for each camera tested. The overall score ranges from 0 to 100, with higher scores indicating better overall test results. Selecting a camera with many options can be a difficult process, and price is certainly a key issue for most consumers. By spending more, will a consumer really get a superior camera? And, do cameras that have more megapixels, a factor often considered to be a good measure of picture quality, cost more than cameras with fewer megapixels? Table 14.15 shows the brand, average retail

TABLE 14.15 DATA FOR 28 POINT-AND-SHOOT DIGITAL CAMERAS

Observation	Brand	Price (\$)	Megapixels	Weight (oz.)	Score
1	Canon	330	10	7	66
2	Canon	200	12	5	66
3	Canon	300	12	7	65
4	Canon	200		6	62
5	Canon	180			62
6	Canon	200			61
7	Canon	200			60
8	Canon	130	14		60
9	Canon	130	10		60
10	Canon	110	12	5	59
11	Canon	90	16	5	55
12	Canon	100	14	5	52
13	Canon		10	6	51

Coefficient of Variation

In some situations we may be interested in a descriptive statistic that indicates how large the standard deviation is relative to the mean. This measure is called the **coefficient of variation** and is usually expressed as a percentage.

COEFFICIENT OF VARIATION

$$\left(\frac{\text{Standard deviation}}{\text{Mean}} \times 100 \right) \% \quad (3.11)$$

For the class size data, we found a sample mean of 44 and a sample standard deviation of 8. The coefficient of variation is $[(8/44) \times 100]\% = 18.2\%$. In words, the coefficient of variation tells us that the sample standard deviation is 18.2% of the value of the sample mean. For the starting salary data with a sample mean of 3940 and a sample standard deviation of 165.65, the coefficient of variation, $[(165.65/3940) \times 100]\% = 4.2\%$, tells us the sample standard deviation is only 4.2% of the value of the sample mean. In general, the coefficient of variation is a useful statistic for comparing the variability of variables that have different standard deviations and different means.

Using Excel's Descriptive Statistics Tool

As we have seen, Excel provides statistical functions to compute descriptive statistics for a data set. These functions can be used to compute one statistic at a time (e.g., mean, variance, etc.). Excel also provides a variety of data analysis tools. One of these, called Descriptive Statistics, allows the user to compute a variety of descriptive statistics at once. We will now show how Excel's Descriptive Statistics tool can be used for the starting salary data in Table 3.1. Refer to Figure 3.7 as we describe the tasks involved.

Enter/Access Data: Open the DATAfile named *StartSalary*. The data are in cells B2:B13 and labels appear in column A and in cell B1.

Apply Tools: The following steps describe how to use Excel's Descriptive Statistics tool for these data.

Step 1. Click the **Data** tab on the Ribbon

Step 2. In the **Analysis** group, click **Data Analysis**

Step 3. Choose **Descriptive Statistics** from the list of **Analysis Tools**

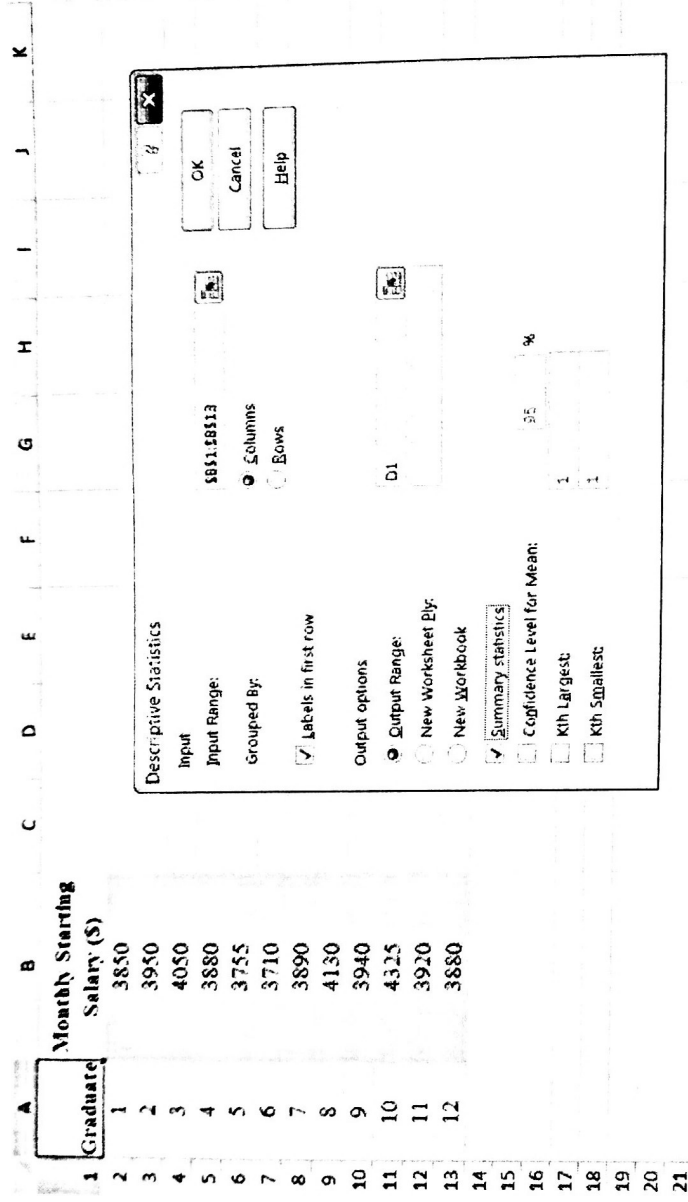
Step 4. When the Descriptive Statistics dialog box appears (see Figure 3.7):
Enter B1:B13 in the **Input Range** box
Select **Grouped By Columns**

Select **Labels in First Row**

Select **Output Range**
Enter D1 in the **Output Range** box (to identify the upper left corner of

the section of the worksheet where the descriptive statistics will appear)
Select **Summary Statistics**
Click **OK**

FIGURE 3.7 DIALOG BOX FOR EXCEL'S DESCRIPTIVE STATISTICS TOOL



Cells D1:D15 of Figure 3.8 show the descriptive statistics provided by Excel. A gold screen is used to highlight the results. The boldfaced entries are the descriptive statistics that we have already covered. The descriptive statistics that are not boldfaced are either covered subsequently in the text or discussed in more advanced texts.

FIGURE 3.8 DESCRIPTIVE STATISTICS PROVIDED BY EXCEL FOR THE STARTING SALARY DATA

Monthly Starting Salary (\$)		Monthly Starting Salary (\$)	
1	Graduate		
2	1	3850	
3	2	3950	
4	3	4050	
5	4	3880	
6	5	3755	
7	6	3710	
8	7	3890	
9	8	4130	
10	9	3940	
11	10	4325	
12	11	3920	
13	12	3880	
14			
15			
16			
		Mean	3940
		Standard Error	47.8199
		Median	3905
		Mode	3880
		Standard Deviation	165.65
		Sample Variance	27440.91
		Kurtosis	1.72
		Skewness	1.09
		Range	615
		Minimum	3710
		Maximum	4325
		Sum	47280
		Count	12

DATA

RaceHelmets

Helmet	Weight (oz)	Price (\$)
Pyrotec Pro Airflow	64	248
Pyrotec Pro Airflow Graphics	64	278
RCi Full Face	64	200
RaceQuip Ridgelinc	64	200
HJC AR-10	58	300
HJC Si-12	47	700
HJC HX-10	49	900
Impact Racing Super Sport	59	340
Zamp FSA-1	66	199
Zamp RZ-2	58	299
Zamp RZ-2 Ferrari	58	299
Zamp RZ-3 Sport	52	479
Zamp RZ-3 Sport Painted	52	479
Bell M2	63	369
Bell M4	62	369
Bell M4 Pro	54	559
G Force Pro Force 1	63	250
G Force Pro Force 1 Grafx	63	280

- Develop a scatter diagram with weight as the independent variable.
- Does there appear to be any relationship between these two variables?
- Develop the estimated regression equation that could be used to predict the price given the weight.
- Test for the significance of the relationship at the .05 level of significance.
- Did the estimated regression equation provide a good fit? Explain.

14.8

Residual Analysis: Validating Model Assumptions

Residual analysis is the primary tool for determining whether the assumed regression model is appropriate.

As we noted previously, the *residual* for observation i is the difference between the observed value of the dependent variable (y_i) and the predicted value of the dependent variable ($\hat{y}_i$).

RESIDUAL FOR OBSERVATION i

$$y_i - \hat{y}_i$$

(14.28)

where

y_i is the observed value of the dependent variable

$\hat{y}_i$ is the predicted value of the dependent variable

In other words, the i th residual is the error resulting from using the estimated regression equation to predict the value of the dependent variable. The residuals for the Armand's Pizza Parlors example are computed in Table 14.7. The observed values of the dependent variable are in the second column and the predicted values of the dependent variable, obtained using the estimated regression equation $\hat{y} = 60 + 5x$, are in the third column. An analysis of the corresponding residuals in the fourth column will help determine whether the assumptions made about the regression model are appropriate.

TABLE 14.7 RESIDUALS FOR ARMAND'S PIZZA PARLORS

Student Population	Sales	Predicted Sales	Residuals
x_i	y_i	$\hat{y}_i = 60 + 5x_i$	$y_i - \hat{y}_i$
2	58	70	-12
6	105	90	15
8	88	100	-12
8	118	100	18
12	117	120	-3
16	137	140	-3
20	157	160	-3
20	169	160	9
22	149	170	-21
26	202	190	12

Let us now review the regression assumptions for the Armand's Pizza Parlors example. A simple linear regression model was assumed.

$$y = \beta_0 + \beta_1 x + \epsilon \quad (14.29)$$

This model indicates that we assumed quarterly sales (y) to be a linear function of the size of the student population (x) plus an error term ϵ . In Section 14.4 we made the following assumptions about the error term ϵ .

1. $E(\epsilon) = 0$.
2. The variance of ϵ , denoted by σ^2 , is the same for all values of x .
3. The values of ϵ are independent.
4. The error term ϵ has a normal distribution.

These assumptions provide the theoretical basis for the t test and the F test used to determine whether the relationship between x and y is significant, and for the confidence and prediction interval estimates presented in Section 14.6. If the assumptions about the error term ϵ appear questionable, the hypothesis tests about the significance of the regression relationship and the interval estimation results may not be valid.

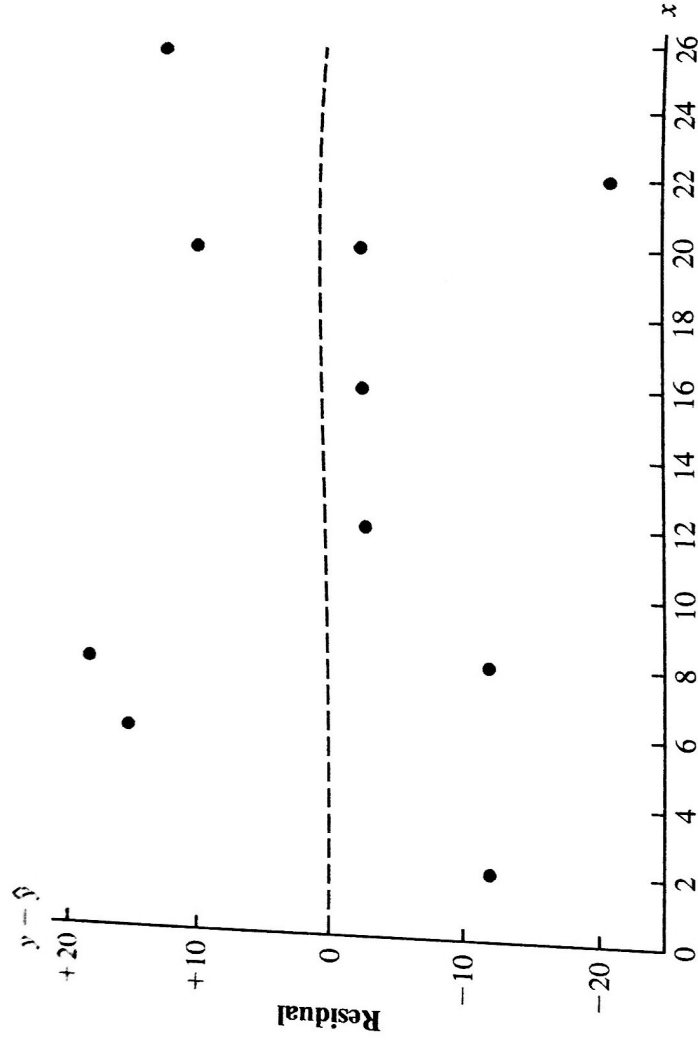
The residuals provide the best information about ϵ ; hence an analysis of the residuals is an important step in determining whether the assumptions for ϵ are appropriate. Much of residual analysis is based on an examination of graphical plots. In this section, we discuss the following residual plots.

1. A plot of the residuals against values of the independent variable x
2. A plot of residuals against the predicted values of the dependent variable y
3. A standardized residual plot
4. A normal probability plot

Residual Plot Against x

A residual plot against the independent variable x is a graph in which the values of the independent variable are represented by the horizontal axis and the corresponding residual values are represented by the vertical axis. A point is plotted for each residual. The first coordinate for each point is given by the value of x_i and the second coordinate is given by the corresponding value of the residual $y_i - \hat{y}_i$. For a residual plot against x with the Armand's Pizza Parlors data from Table 14.7, the coordinates of the first point are (2, -12), corresponding to $x_1 = 2$ and $y_1 - \hat{y}_1 = -12$; the coordinates of the second point are (6, 15), corresponding to $x_2 = 6$ and $y_2 - \hat{y}_2 = 15$; and so on. Figure 14.15 shows the resulting residual plot.

FIGURE 14.15 PLOT OF THE RESIDUALS AGAINST THE INDEPENDENT VARIABLE x
FOR ARMAND'S PIZZA PARLORS



Before interpreting the results for this residual plot, let us consider some general patterns that might be observed in any residual plot. Three examples appear in Figure 14.16. If the assumption that the variance of ϵ is the same for all values of x and the assumed regression model is an adequate representation of the relationship between the variables, the residual plot should give an overall impression of a horizontal band of points such as the one in Panel A of Figure 14.16. However, if the variance of ϵ is not the same for all values of x —for example, if variability about the regression line is greater for larger values of x —a pattern such as the one in Panel B of Figure 14.16 could be observed. In this case, the assumption of a constant variance of ϵ is violated. Another possible residual plot is shown in Panel C. In this case, we would conclude that the assumed regression model is not an adequate representation of the relationship between the variables. A curvilinear regression model or multiple regression model should be considered.

Now let us return to the residual plot for Armand's Pizza Parlors shown in Figure 14.15. The residuals appear to approximate the horizontal pattern in Panel A of Figure 14.16. Hence, we conclude that the residual plot does not provide evidence that the assumptions made for Armand's regression model should be challenged. At this point, we are confident in the conclusion that Armand's simple linear regression model is valid.

Experience and good judgment are always factors in the effective interpretation of residual plots. Seldom does a residual plot conform precisely to one of the patterns in Figure 14.16. Yet analysts who frequently conduct regression studies and frequently review residual plots become adept at understanding the differences between patterns that are reasonable and patterns that indicate the assumptions of the model should be questioned. A residual plot provides one technique to assess the validity of the assumptions for a regression model.

Residual Plot Against $\hat{y}$

Another residual plot represents the predicted value of the dependent variable $\hat{y}$ on the horizontal axis and the residual $y - \hat{y}$ on the vertical axis. This plot is useful for checking the assumptions of the regression model.

FIGURE 14.16 RESIDUAL PLOTS FROM THREE REGRESSION STUDIES

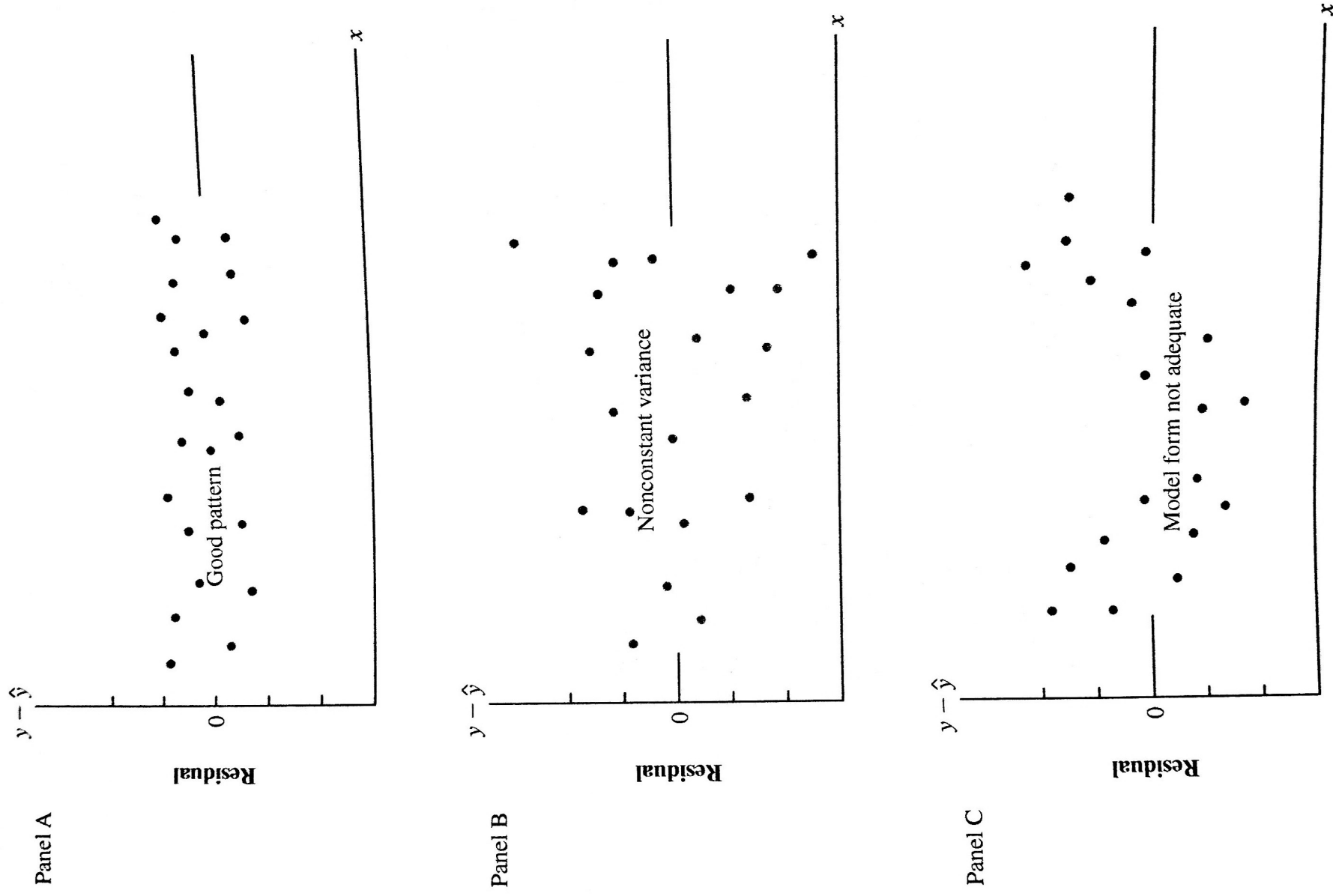
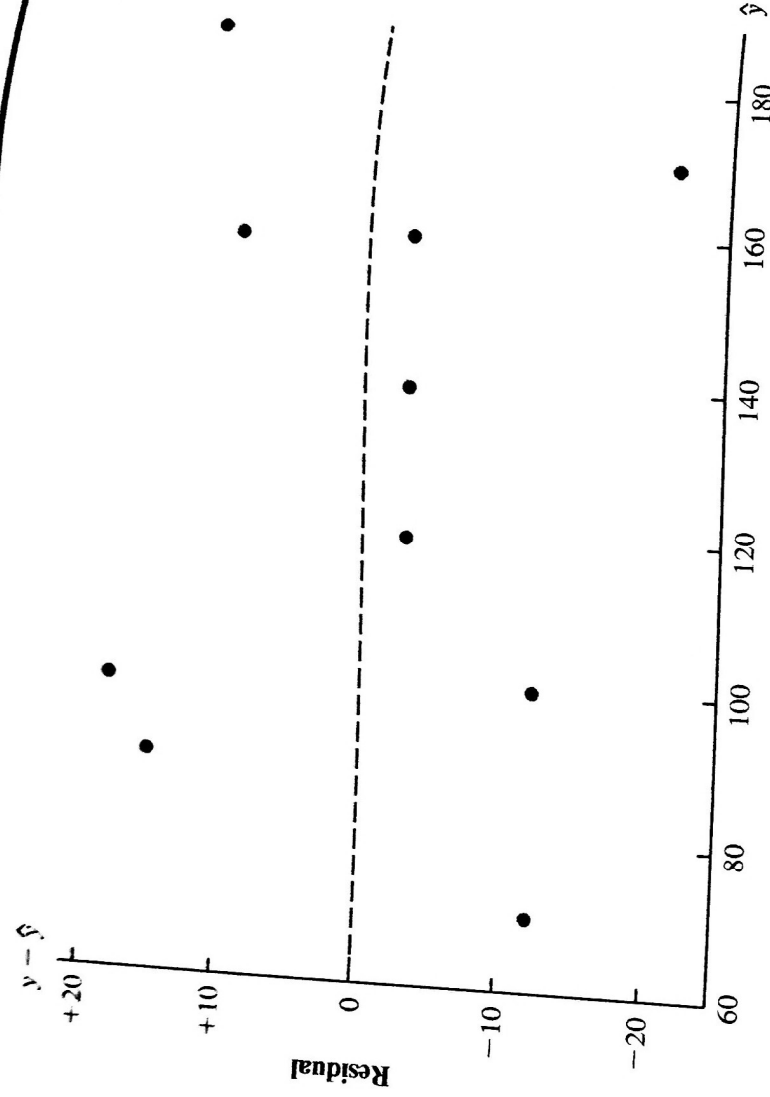


FIGURE 14.17 PLOT OF THE RESIDUALS AGAINST THE PREDICTED VALUES $\hat{y}$
FOR ARMAND'S PIZZA PARLORS



The first coordinate for each point is given by $\hat{y}_i$ and the second coordinate is given by the corresponding value of the i th residual $y_i - \hat{y}_i$. With the Armand's data from Table 14.7, the coordinates of the first point are $(70, -12)$, corresponding to $\hat{y}_1 = 70$ and $y_1 - \hat{y}_1 = -12$; the coordinates of the second point are $(90, 15)$; and so on. Figure 14.17 provides the residual plot. Note that the pattern of this residual plot is the same as the pattern of the residual plot against the independent variable x . It is not a pattern that would lead us to question the model assumptions. For simple linear regression, both the residual plot against x and the residual plot against $\hat{y}$ provide the same pattern. For multiple regression analysis, the residual plot against $\hat{y}$ is more widely used because of the presence of more than one independent variable.

Standardized Residuals

Many of the residual plots provided by computer software packages use a standardized version of the residuals. As demonstrated in preceding chapters, a random variable is standardized by subtracting its mean and dividing the result by its standard deviation. With the least squares method, the mean of the residuals is zero. Thus, simply dividing each residual by its standard deviation provides the **standardized residual**.

It can be shown that the standard deviation of residual i depends on the standard error of the estimate s and the corresponding value of the independent variable x_i ,

STANDARD DEVIATION OF THE i th RESIDUAL²

$$s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i} \quad (14.30)$$

²This equation actually...

14.8 Residual Analysis: Validating Model Assumptions

where

$s_{y_i - \hat{y}_i}$ = the standard deviation of residual i

s = the standard error of the estimate

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2} \quad (14.31)$$

Note that equation (14.30) shows that the standard deviation of the i th residual depends on x_i because of the presence of h_i in the formula.³ Once the standard deviation of each residual is calculated, we can compute the standardized residual by dividing each residual by its corresponding standard deviation.

STANDARDIZED RESIDUAL FOR OBSERVATION i

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}} \quad (14.32)$$

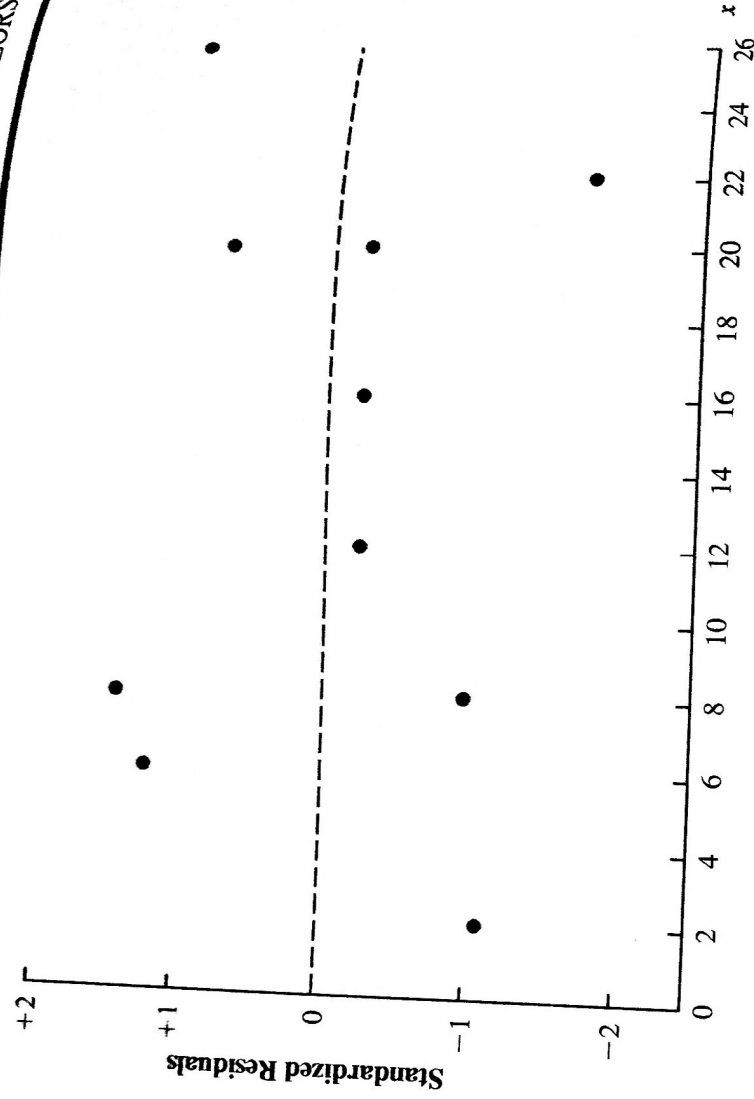
Table 14.8 shows the calculation of the standardized residuals for Armand's Pizza Parlors. Recall that previous calculations showed $s = 13.829$. Figure 14.18 is the plot of the standardized residuals against the independent variable x .

TABLE 14.8 COMPUTATION OF STANDARDIZED RESIDUALS FOR ARMAND'S PIZZA PARLORS

Restaurant i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$	$\frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}$	h_i	$s_{y_i - \hat{y}_i}$	$y_i - \hat{y}_i$	Standardized Residual
1	2	-12	144	.2535	.3535	11.1193	-12	-1.0792
2	6	-8	64	.1127	.2127	12.2709	15	1.2224
3	8	-6	36	.0634	.1634	12.6493	-12	-.9487
4	8	-6	36	.0634	.1634	12.6493	18	1.4230
5	12	-2	4	.0070	.1070	13.0682	-3	-.2296
6	16	2	4	.0070	.1070	13.0682	-3	-.2296
7	20	6	36	.0634	.1634	12.6493	-3	-.2372
8	20	6	36	.0634	.1634	12.6493	9	.7115
9	22	8	64	.1127	.2127	12.2709	-21	-1.7114
10	26	12	144	.2535	.3535	11.1193	12	1.0792
Total								568

Note: The values of the residuals were computed in Table 14.7.

FIGURE 14.18 PLOT OF THE STANDARDIZED RESIDUALS AGAINST THE INDEPENDENT VARIABLE x FOR ARMAND'S PIZZA PARLORS



Small departures from normality do not have a great effect on the statistical tests used in regression analysis.

The standardized residual plot can provide insight about the assumption that the error term ϵ has a normal distribution. If this assumption is satisfied, the distribution of the standardized residuals should appear to come from a standard normal probability distribution.⁴ Thus, when looking at a standardized residual plot, we should expect to see approximately 95% of the standardized residuals between -2 and $+2$. We see in Figure 14.18 that for the Armand's example all standardized residuals are between -2 and $+2$. Therefore, on the basis of the standardized residuals, this plot gives us no reason to question the assumption that ϵ has a normal distribution.

Because of the effort required to compute the estimated values of $\hat{y}$, the residuals, and the standardized residuals, most statistical packages provide these values as optional regression output. Hence, residual plots can be easily obtained. For large problems computer packages are the only practical means for developing the residual plots discussed in this section.

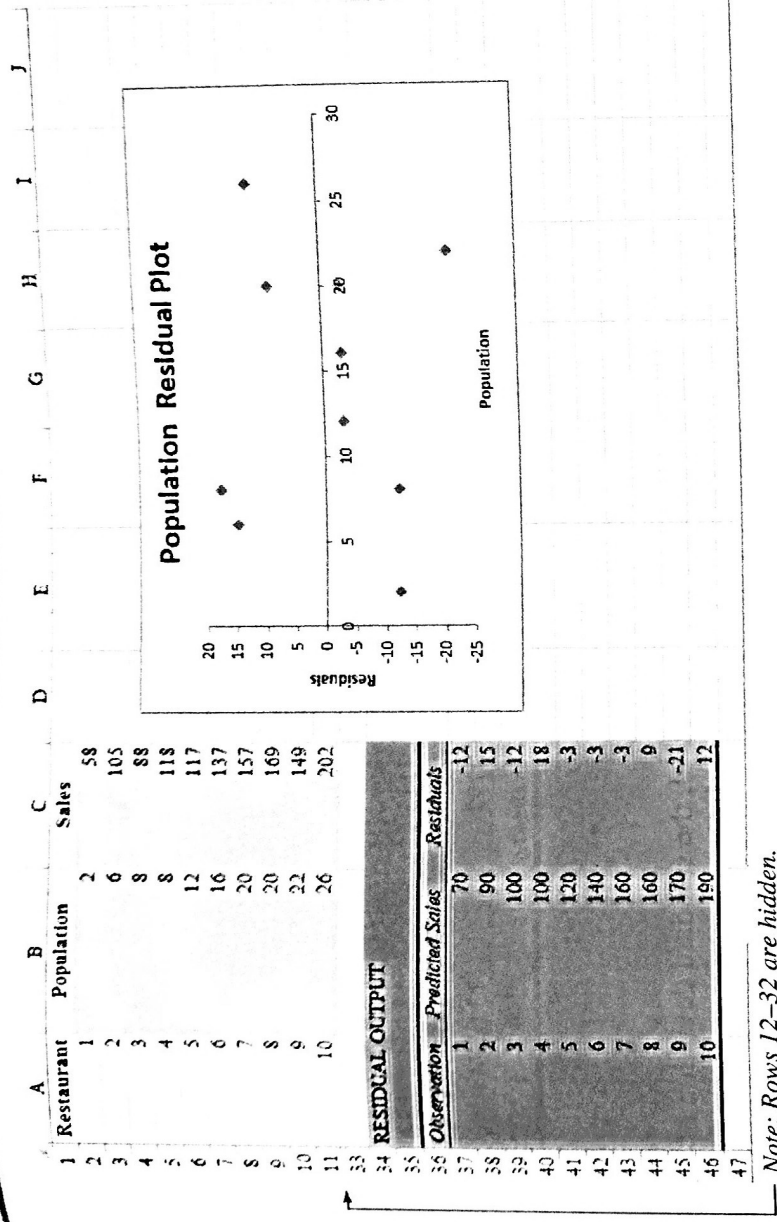
Using Excel to Construct a Residual Plot

In Section 14.7 we showed how Excel's Regression tool could be used for regression analysis. The Regression tool also provides the capability to obtain a residual plot against the independent variable x and, when used with Excel's chart tools, the Regression tool residual output can also be used to construct a residual plot against $\hat{y}$ as well as an Excel version of a standardized residual plot.

Residual plot against x To obtain a residual plot against x , the steps that we describe in Section 14.7 in order to obtain the regression output are performed with one change.

⁴Because s is used instead of σ in equation (14.30), the probability distribution of the standardized residuals is not technically normal. However, in most regression studies, the sample size is large enough that a normal approximation is very good.

FIGURE 14.19 REGRESSION TOOL RESIDUAL OUTPUT FOR THE ARMAND'S PIZZA PARLORS PROBLEM



When the Regression tool dialog box appears (see Figure 14.13), we must also select the Residual Plots option in the Residual section. The regression output will appear as described previously, and the worksheet will also contain a chart showing a plot of the residuals against the independent variable Population. In addition, a list of predicted values of y and the corresponding residual values are provided below the regression output. Figure 14.19 shows the residual output for the Armand's Pizza Parlors problem; note that rows 12–32, containing the standard Regression tool output, have been hidden to better focus on the residual portion of the output. We see that the shape of this plot is the same as shown previously in Figure 14.15.

Residual plot against $\hat{y}$ Using Excel's chart tools and the residual output provided in Figure 14.19, we can easily construct a residual plot against $\hat{y}$. The following steps describe how to use Excel's chart tools to construct the residual plot using the regression tool output in the worksheet.

Step 1. Select cells B37:C46

Step 2. Click the **Insert** tab on the Ribbon

Step 3. In the **Charts** group, click **Insert Scatter (X, Y) or Bubble Chart**

Step 4. When the list of scatter diagram subtypes appears:

Click **Scatter with only Markers** (the chart in the upper left corner)

The resulting chart will look similar to the residual plot shown in Figure 14.17. Adding a chart title, labels for the horizontal and vertical axes, as well as other formatting options can be easily done. Note that except for using different data to construct the chart as well as different labels for the chart output, the steps describing how to use Excel's chart tools to construct a residual plot against $\hat{y}$ are the same as the steps we used to construct a scatter diagram in Section 14.2.

Excel's standardized residual plot Excel can be used to construct what it calls a standardized residual plot. Excel's standardized residual plot is really an approximation

TABLE 14.9 COMPUTATION OF EXCEL'S STANDARD RESIDUALS

Restaurant i	Values from Table 14.8		Values Using Excel	
	$y_i - \hat{y}_i$	$s_{y_i} - \hat{s}_{y_i}$	Standardized Residual	Estimate of Standard Residual
1	-12	11.1193	-1.0792	13.0384
2	15	12.2709	1.2224	13.0384
3	-12	12.6493	-.9487	13.0384
4	18	12.6493	1.4230	13.0384
5	-3	13.0682	-.2296	13.0384
6	-3	13.0682	-.2296	13.0384
7	-3	12.6493	-.2372	13.0384
8	9	12.6493	.7115	13.0384
9	-21	12.2709	-1.7114	13.0384
10	12	11.1193	1.0792	13.0384
				-0.9204
				1.1504
				-0.9204
				1.3805
				-0.2301
				-0.2301
				-0.2301
				0.6903
				-1.6106
				0.9204

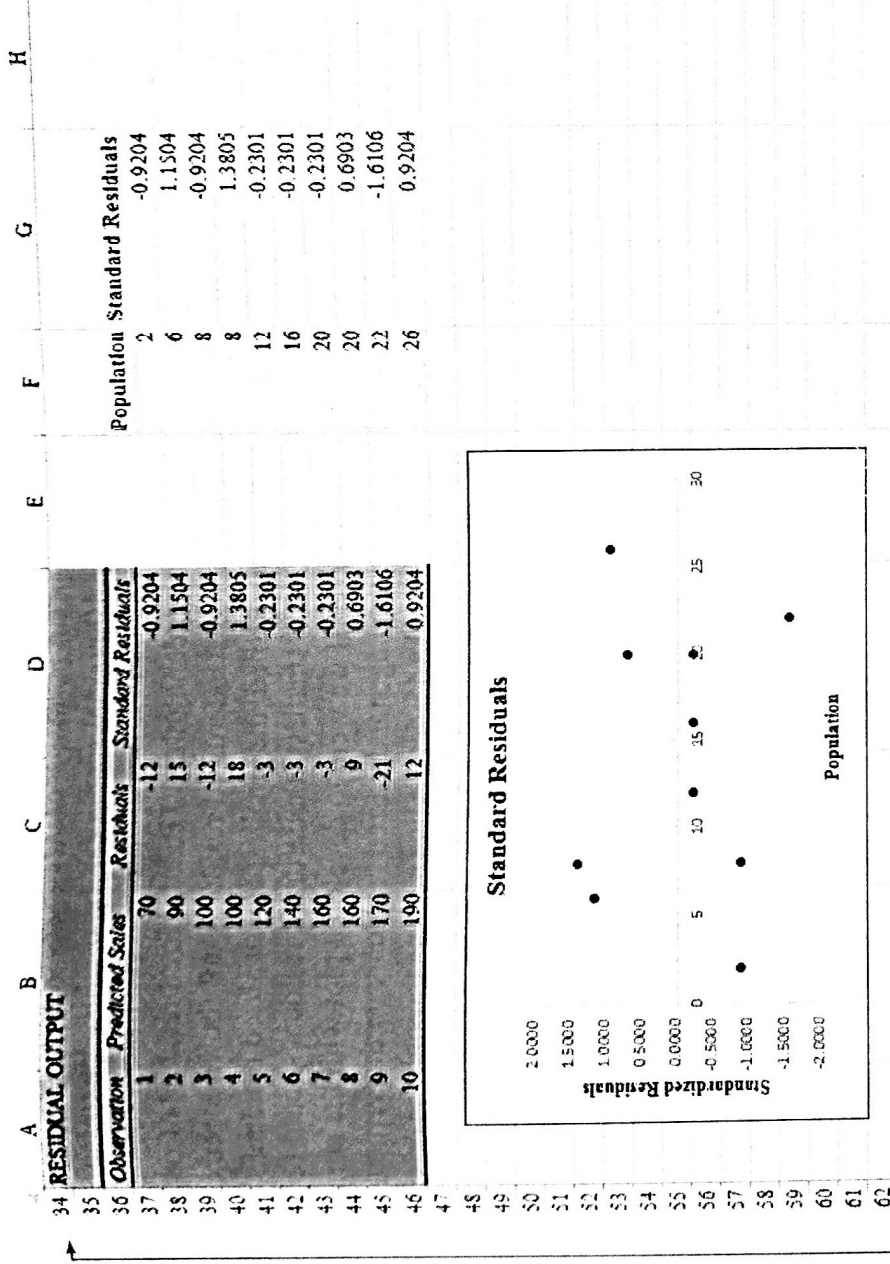
of the true standardized residual plot. In Excel, the standard deviation of the i th residual is not computed using equation (14.30). Instead, Excel estimates $s_{y_i} - \hat{s}_{y_i}$ using the standard deviation of the n residual values. Then, dividing each residual by this estimate, Excel obtains what it refers to as a standard residual. The plot of these standard residuals is what you get when you request a plot of the standardized residuals using Excel's Regression tool.

We will illustrate how to construct a standardized residual plot using Excel for the Armand's Pizza Parlors problem. The residuals for the Armand's Pizza Parlors problem are -12, 15, -12, 18, -3, -3, 9, -21, and 12. Using Excel's STDEV.S function, we computed a standard deviation of 13.0384 for these 10 data values. To compute the standard residuals, Excel divides each residual by 13.0384; the results are shown in Table 14.9. Both the standardized residuals, computed in Table 14.8, and Excel's standard residuals are shown. There is not a great deal of difference between Excel's standard residuals and the true standardized residuals. In general, the differences get smaller as the sample size increases. Often we are interested only in identifying the general pattern of the points in a standardized residual plot; in such cases, the small differences between the standardized residuals and Excel's standard residuals will have little effect on the pattern observed. Thus these differences will not influence the conclusions reached when we use the residual plot to validate model assumptions.

The Regression tool and the chart tools can be used to obtain Excel's standardized residual plot. First, the steps that we described in Section 14.7 in order to conduct a regression analysis are performed with one change. When the Regression dialog box appears (see Figure 14.13), we must select the Standardized Residuals option. In addition to the regression output described previously, the output will contain a list of predicted values of y , residuals, and standard residuals, as shown in cells A34:D46 in Figure 14.20.

The Standardized Residuals option does not automatically produce a standardized residual plot. But we can use Excel's chart tools to construct a scatter diagram in which the values of the independent variable are placed on the horizontal axis and the values of the standard residuals are placed on the vertical axis. The procedure that describes how to use Excel's chart tools to construct a standardized residual plot is similar to the steps we showed for using Excel's chart tools to construct a residual plot against $\hat{y}$. Because Excel requires

14.20 STANDARDIZED RESIDUAL PLOT AGAINST THE INDEPENDENT VARIABLE POPULATION FOR THE ARMAND'S PIZZA PARLORS EXAMPLE



Note: Rows 1–33 are hidden.

that the two variables being plotted be located in adjacent columns of the worksheet, we copied the data and heading for the independent variable Population into cells F36:F46 and the data and heading for the standard residuals into cells G36:G46.

Using the data in cells F36:G46 and Excel's chart tools we obtained the scatter diagram shown in Figure 14.20; this scatter diagram is Excel's version of the standardized residual plot for the Armand's Pizza Parlors example. Comparing Excel's version of the standardized residual plot to the standardized residual plot in Figure 14.17, we see the same pattern evident. All of the standardized residuals in both figures are between -2 and $+2$, indicating no reason to question the assumption that ϵ has a normal distribution.

Normal Probability Plot

Another approach for determining the validity of the assumption that the error term has a normal distribution is the **normal probability plot**. To show how a normal probability plot is developed, we introduce the concept of *normal scores*.

Suppose 10 values are selected randomly from a normal probability distribution with a mean of zero and a standard deviation of one, and that the sampling process is repeated over and over with the values in each sample of 10 ordered from smallest to largest. For now, let us consider only the smallest value in each sample. The random variable representing the smallest value obtained in repeated sampling is called the first-order statistic.

TABLE 14.10
NORMAL SCORES
FOR $n = 10$

Order Statistic	Normal Score
1	-1.55
2	-1.00
3	-.65
4	-.37
5	-.12
6	.12
7	.37
8	.65
9	1.00
10	1.55

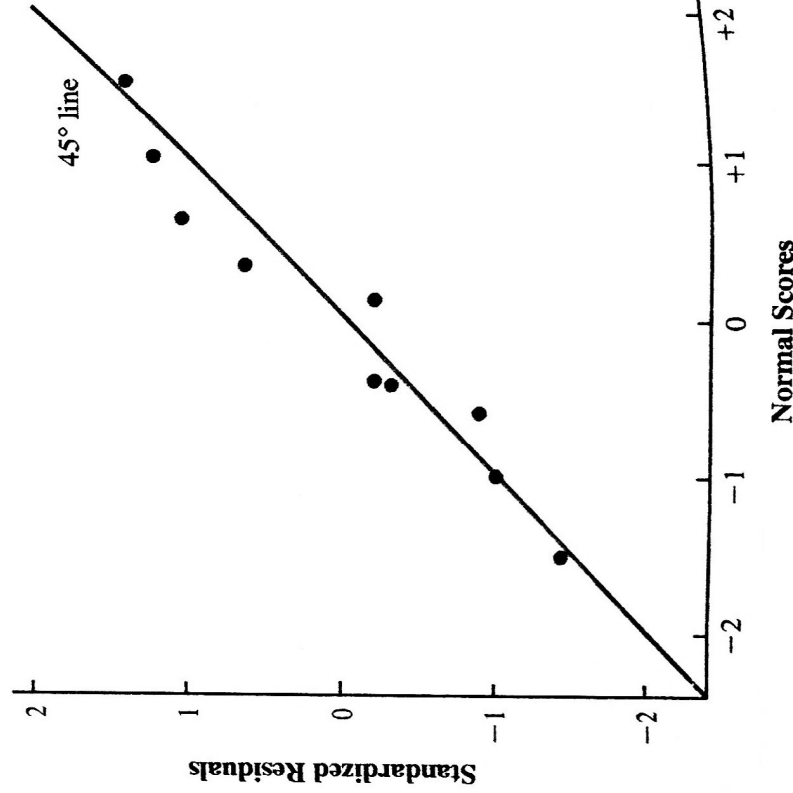
TABLE 14.11
NORMAL SCORES
AND ORDERED
STANDARDIZED
RESIDUALS FOR
ARMAND'S PIZZA
PARLORS

Normal Scores	Ordered Standardized Residuals
-1.55	-1.7114
-1.00	-1.0792
-.65	-.9487
-.37	-.2372
-.12	-.2296
.12	-.2296
.37	.7115
.65	1.0792
1.00	1.2224
1.55	1.4230

Statisticians show that for samples of size 10 from a standard normal probability distribution, the expected value of the first-order statistic is -1.55 . This expected value is called a normal score. For the case with a sample of size $n = 10$, there are 10 order statistics and 10 normal scores (see Table 14.10). In general, a data set consisting of n observations will have n order statistics and hence n normal scores.

Let us now show how the 10 normal scores can be used to determine whether the standardized residuals for Armand's Pizza Parlors appear to come from a standard normal probability distribution. We begin by ordering the 10 standardized residuals from Table 14.8. The 10 normal scores and the ordered standardized residuals are shown together in Table 14.11. If the normality assumption is satisfied, the smallest standardized residual should be close to the smallest normal score, the next smallest standardized residual should be close to the next smallest normal score, and so on. If we were to develop a plot with the normal scores on the horizontal axis and the corresponding standardized residuals on the vertical axis, the plotted points should cluster closely around a 45-degree line passing through the origin if the standardized residuals are approximately normally distributed. Such a plot is referred to as a *normal probability plot*.

Figure 14.21 is the normal probability plot for the Armand's Pizza Parlors example. Judgment is used to determine whether the pattern observed deviates from the line enough to conclude that the standardized residuals are not from a standard normal probability distribution. In Figure 14.21, we see that the points are grouped closely about the line. We therefore conclude that the assumption of the error term having a normal probability distribution is reasonable. In general, the more closely the points are clustered about the 45-degree line, the stronger the evidence supporting the normality assumption. Any substantial curvature in the normal probability plot is evidence that the residuals have not come from a normal distribution.

FIGURE 14.21 NORMAL PROBABILITY PLOT FOR ARMAND'S PIZZA PARLORS

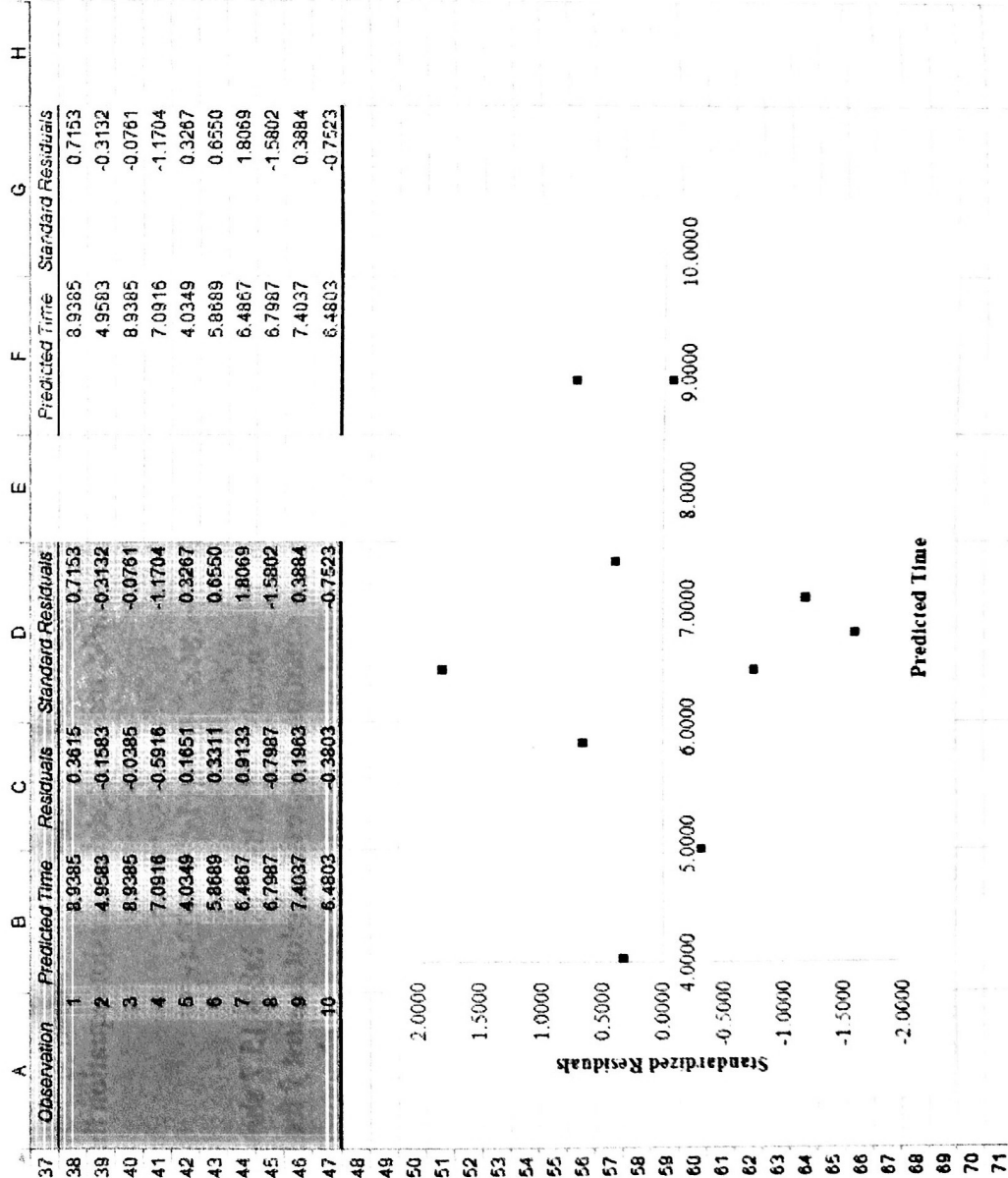
A residual plot against $\hat{y}$ can also be used when performing residual analysis in simple linear regression. In fact, the pattern for a residual plot against $\hat{y}$ in simple linear regression is the same as the pattern of the residual plot against x . Thus, for simple linear regression the residual plot against $\hat{y}$ and the residual plot against x provide the same information. In multiple regression analysis, however, it is preferable to use the residual plot against $\hat{y}$ to determine whether the model's assumptions are satisfied.

Standardized Residual Plot Against $\hat{y}$

In Chapter 14 we pointed out that standardized residuals were frequently used in residual plots. We showed how to construct a standardized residual plot against x and discussed how the standardized residual plot could be used to identify outliers and provide insight about the assumption that the error term ϵ has a normal distribution. Recall that we recommended considering any observation with a standardized residual of less than -2 or greater than $+2$ as an outlier. With normally distributed errors, standardized residuals should be outside these limits approximately 5% of the time.

In multiple regression analysis, the computation of the standardized residuals is too complex to be done by hand. As we showed in Section 14.8, Excel's Regression tool can

FIGURE 15.12 STANDARDIZED RESIDUAL PLOT AGAINST THE PREDICTED VALUES $\hat{y}$
FOR BUTLER TRUCKING



Note: Rows 1–34 are hidden.

be used to compute the standard residuals. In multiple regression analysis we use the same procedure to compute the standard residuals. Instead of developing a standardized residual plot against each of the independent variables, we will construct one standardized residual plot against the predicted values $\hat{y}$.

Figure 15.12 shows the standard residuals and the corresponding standardized residual against $\hat{y}$ (Predicted Time) for Butler Trucking developed using Excel's Regression tool and Excel's chart tools. The standardized residual plot does not indicate any abnormalities, and no standard residual is less than -2 or greater than $+2$. Note that the pattern of the standardized residual plot against $\hat{y}$ is the same as the pattern of the residual plot against $\hat{y}$ shown in Figure 15.11. But the standardized residual plot is preferred because it enables us to check for outliers and determine whether the assumption of normality for the regression model is reasonable.



Methods

39. Data for two variables, x and y , follow.

x_i	1	2	3	4	5
y_i	3	7	5	11	14

- Develop the estimated regression equation for these data.
 - Plot the residuals against $\hat{y}$. Does the residual plot support the assumptions about ϵ ? Explain.
 - Plot the standardized residuals against $\hat{y}$. Do any outliers appear in these data? Explain.
40. Data for two variables, x and y , follow.

x_i	22	24	26	28	40
y_i	12	21	31	35	70

- Develop the estimated regression equation for these data.
- Compute the standardized residuals for these data. Can any of these observations be classified as an outlier? Explain.
- Develop a standardized residual plot against $\hat{y}$. Does the residual plot support the assumptions about ϵ ? Explain.

Applications

41. Exercise 5 gave the following data on weekly gross revenue (\$1000s), television advertising expenditures (\$1000s), and newspaper advertising expenditures (\$1000s) for Showtime Movie Theaters.

Weekly Gross Revenue (\$1000s)	Television Advertising (\$1000s)	Newspaper Advertising (\$1000s)
96	5.0	1.5
90	2.0	2.0
95	4.0	1.5
92	2.5	2.5
95	3.0	3.3
95	3.5	2.3
94	2.5	4.2

Chapter 14 Simple Linear Regression

Chapter 14 Simple Linear Regression
 associated with $t = 8.62$ must be less than $2(.005) = .01$. Using Excel, the p -value = .000. Because the p -value is less than $\alpha = .01$, we reject H_0 and conclude that β_1 is not equal to zero. This evidence is sufficient to conclude that a significant relationship exists between student population and quarterly sales. A summary of the t test for significance in simple linear regression follows.

t TEST FOR SIGNIFICANCE IN SIMPLE LINEAR REGRESSION

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

TEST STATISTIC

$$t = \frac{b_1}{s_{b_1}} \quad (14.19)$$

REJECTION RULE

p -value approach: Reject H_0 if $p\text{-value} \leq \alpha$

Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - 2$ degrees of freedom.

Confidence Interval for β_1

The form of a confidence interval for β_1 is as follows:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

The point estimator is b_1 and the margin of error is $t_{\alpha/2} s_{b_1}$. The confidence coefficient associated with this interval is $1 - \alpha$, and $t_{\alpha/2}$ is the t value providing an area of $\alpha/2$ in the upper tail of a t distribution with $n - 2$ degrees of freedom. For example, suppose that we wanted to develop a 99% confidence interval estimate of β_1 for Armand's Pizza Parlors. From Table 2 of Appendix B we find that the t value corresponding to $\alpha = .01$ and $n - 2 = 10 - 2 = 8$ degrees of freedom is $t_{.005} = 3.355$. Thus, the 99% confidence interval estimate of β_1 is

$$b_1 \pm t_{\alpha/2} s_{b_1} = 5 \pm 3.355(.5803) = 5 \pm 1.95$$

or 3.05 to 6.95.

In using the t test for significance, the hypotheses tested were

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

At the $\alpha = .01$ level of significance, we can use the 99% confidence interval as an alternative for drawing the hypothesis testing conclusion for the Armand's data. Because 0, the hypothesized value of β_1 , is not included in the confidence interval (3.05 to 6.95), we can reject H_0 and conclude that a significant statistical relationship exists between the size of the student population and quarterly sales. In general, a confidence interval can be used to

$$F \text{ TEST FOR SIGNIFICANCE IN SIMPLE LINEAR REGRESSION}$$

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

If H_0 is false, MSE still provides an unbiased estimate of σ^2 and MSR overestimates σ^2 . If H_0 is true, both MSE and MSR provide unbiased estimates of σ^2 ; in this case the value of MSR/MSE should be close to 1.

TEST STATISTIC

$$F = \frac{\text{MSR}}{\text{MSE}} \quad (14.21)$$

REJECTION RULE

p -value approach: Reject H_0 if $p\text{-value} \leq \alpha$
 Critical value approach: Reject H_0 if $F \geq F_\alpha$

where F_α is based on an F distribution with 1 degree of freedom in the numerator and $n - 2$ degrees of freedom in the denominator.

In Chapter 13 we covered analysis of variance (ANOVA) and showed how an ANOVA table could be used to provide a convenient summary of the computational aspects of analysis of variance. A similar ANOVA table can be used to summarize the results of the F test for significance in regression. Table 14.5 is the general form of the ANOVA table for simple linear regression. Table 14.6 is the ANOVA table with the F test computations performed for Armand's Pizza Parlors. Regression, Error, and Total are the labels for the three sources of variation, with SSR, SSE, and SST appearing as the corresponding sum of

TABLE 14.5 GENERAL FORM OF THE ANOVA TABLE FOR SIMPLE LINEAR REGRESSION

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p -Value
Regression	SSR	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$F = \frac{\text{MSR}}{\text{MSE}}$	
Error	SSE	$n - 2$	$\text{MSE} = \frac{\text{SSE}}{n - 2}$		
Total	SST	$n - 1$			

In every analysis of variance table the total sum of squares is the sum of the regression sum of squares and the error sum of squares; in addition, the total degrees of freedom is the sum of the regression degrees of freedom and the error degrees of freedom.

TABLE 14.6 ANOVA TABLE FOR THE ARMAND'S PIZZA PARLORS PROBLEM

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p -Value
Regression	14,200	1	$\frac{14,200}{1} = 14,200$	$\frac{14,200}{191.25} = 74.25$	.000
Error	1,530	8	$\frac{1,530}{8} = 191.25$		
Total	15,730	9			

18. Refer to exercise 10, where Major League Baseball (MLB) pitching statistics were reported for a random sample of 20 pitchers from the American League for one full season.
 - a. In part (c) of exercise 10, an estimated regression equation was developed relating the average number of runs given up per inning pitched given the average number of strikeouts per inning pitched and the average number of home runs per inning pitched. What are the values of R^2 and R_a^2 ?
 - b. Does the estimated regression equation provide a good fit to the data? Explain.
 - c. Suppose the earned run average (ERA) is used as the dependent variable in part (c) instead of the average number of runs given up per inning pitched. Does the estimated regression equation that uses the ERA provide a good fit to the data? Explain.

15.4

Model Assumptions

In Section 15.1 we introduced the following multiple regression model.

MULTIPLE REGRESSION MODEL

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p + \epsilon \quad (15.10)$$

The assumptions about the error term ϵ in the multiple regression model parallel those for the simple linear regression model.

ASSUMPTIONS ABOUT THE ERROR TERM ϵ IN THE MULTIPLE REGRESSION MODEL $y = \beta_0 + \beta_1x_1 + \cdots + \beta_px_p + \epsilon$

1. The error term ϵ is a random variable with mean or expected value of zero; that is, $E(\epsilon) = 0$.
Implication: For given values of $x_1, x_2, \dots, x_p$, the expected, or average, value of y is given by

$$E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \quad (15.11)$$

- Equation (15.11) is the multiple regression equation we introduced in Section 15.1. In this equation, $E(y)$ represents the average of all possible values of y that might occur for the given values of $x_1, x_2, \dots, x_p$.
2. The variance of ϵ is denoted by σ^2 and is the same for all values of the independent variables $x_1, x_2, \dots, x_p$.
Implication: The variance of y about the regression line equals σ^2 and is the same for all values of $x_1, x_2, \dots, x_p$.
3. The values of ϵ are independent.
Implication: The value of ϵ for a particular set of values for the independent variables is not related to the value of ϵ for any other set of values.
4. The error term ϵ is a normally distributed random variable reflecting the deviation between the y value and the expected value of y given by $\beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p$.
Implication: Because $\beta_0, \beta_1, \dots, \beta_p$ are constants for the given values of $x_1, x_2, \dots, x_p$, the dependent variable y is also a normally distributed random variable.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F	p -Value
Regression	SSR	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$	
Error	SSE	$n - p - 1$	$MSE = \frac{SSE}{n - p - 1}$		
Total	SST	$n - 1$			

Table 15.3 is the general form of the ANOVA table for multiple regression. The value of the F test statistic and its corresponding p -value in the last column can be used to make the hypothesis test conclusion. By reviewing the Excel output for Butler Trucking Company in Figure 15.7, we see that Excel’s analysis of variance table contains this information.

t Test

If the F test shows that the multiple regression relationship is significant, a t test can be conducted to determine the significance of each of the individual parameters. The t test for individual significance follows.

t TEST FOR INDIVIDUAL SIGNIFICANCE

For any parameter β_i

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$

TEST STATISTIC

$$t = \frac{b_i}{s_{b_i}} \tag{15.15}$$

REJECTION RULE

p -value approach: Reject H_0 if $p\text{-value} \leq \alpha$

Critical value approach: Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with $n - p - 1$ degrees of freedom.

In the test statistic, s_{b_i} is the estimate of the standard deviation of b_i . The value of s_{b_i} will be provided by the computer software package.

Let us conduct the t test for the Butler Trucking regression problem. Refer to the section of Figure 15.7 that shows the Excel output for the t -ratio calculations. Values of b_1 , b_2 , s_{b_1} , and s_{b_2} are as follows.

$$b_1 = .0611 \qquad s_{b_1} = .0099$$

$$b_2 = .9234 \qquad s_{b_2} = .2211$$