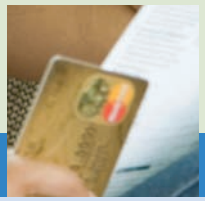
CHAPTER 14

Analyzing Linear Relationships, Two or More Variables



INTRODUCTION

In the previous chapter, we introduced Kate Cameron, the owner of Woodbon, a company that produces high-quality wooden furniture. Kate wanted to understand why sales have grown steadily over recent years, with an eye to planning for the future. After carefully checking required conditions for the analysis, Kate created a mathematical model of the relationship between Woodbon's sales and advertising. While there did seem to be a significant relationship between the two variables, the variability in the predictions meant that the model was not that useful for predicting sales.

The discussion in the previous chapter was a useful introduction to analyzing relationships between two variables. However, a more realistic process would begin with Kate analyzing the relationship between Woodbon's sales and a number of possible explanatory variables. Some that have already been mentioned are housing starts and mortgage rates. Kate would likely examine a number of possible explanatory variables, with the aim of developing a model that is economical (that is, has reasonable data requirements) and works well (that is, makes useful predictions).

LEARNING OBJECTIVES

After mastering the material in this chapter, you will be able to:

1. Estimate the linear relationship between a quantitative response variable and one or more explanatory variables.
2. Check the conditions required for use of the regression model in hypothesis testing and prediction.
3. Assess the regression relationship, using appropriate hypothesis tests and a coefficient of determination.
4. Make predictions using the regression relationship.
5. Understand the considerations involved in choosing the "best" regression model, and the challenges presented by multicollinearity.
6. Use indicator variables to model qualitative explanatory variables.

Section 14.1 builds on the discussion in Chapter 13, to extend the mathematical model to include more than one explanatory variable. A reasonable way to start is with some careful thinking about what other factors could most reasonably be expected to affect Woodbon's sales. For more complex models, it is crucial to have computer software to do the calculations, and you will see how to use Excel to build the mathematical model.

Section 14.2 extends the theoretical model from the last chapter to include more explanatory variables, revisiting the discussion about least-squares models. As before, we will use Excel to check the required conditions for the regression model.

Section 14.3 introduces hypothesis tests about the significance of the overall model, and the individual explanatory variables. We will also discuss a measure of the strength of the relationship between the explanatory variables and the response variables, the adjusted coefficient of determination (adjusted R^2).

Section 14.4 describes an Excel add-in that you can use to make predictions of average and individual response variables, given specific values of the explanatory variables in the model.

In Section 14.5, we will discuss an approach to selecting the best explanatory variables for our regression model. Kate will want to develop a model of sales that makes good predictions, but the simplest model that does a good job will be preferred. Selecting the appropriate explanatory variables is an art as well as a science. An Excel add-in that produces a summary of all possible models will be introduced.

In Section 14.5, we will look at ways to assess and deal with a new problem that may arise when there is more than one explanatory variable. This problem is usually referred to as "multicollinearity," and it occurs when one of the explanatory variables is related to one or more of the other explanatory variables.

It is possible to include qualitative explanatory variables in regression models, and Section 14.6 illustrates the use of indicator variables to accomplish this.

Section 14.7 refers briefly to more advanced models, so that you can get a sense of the wide variety of mathematical modelling possibilities.

14.1 DETERMINING THE RELATIONSHIP—MULTIPLE LINEAR REGRESSION

In Chapter 13, Kate Cameron examined the relationship between advertising spending and sales. This simple linear regression model served as an introduction to the techniques of linear regression modelling.

It seems reasonable to think that there is a cause-and-effect relationship between advertising and sales. Kate is also wondering if sales are significantly affected by other explanatory variables. In particular, she is wondering about three others:

- Mortgage rates may affect a household's ability to buy furniture. Kate expects the relationship to be negative, that is, when mortgage rates are higher, she would expect a household to have less income available to buy wooden furniture.

- Housing starts may also be related to sales. When more houses are being built, more households might buy Woodbon's furniture. There is a fairly long lead time for a customer to take delivery of Woodbon's furniture, so housing starts may be a useful explanatory variable.
- Kate has been exploring Statistics Canada data, and has discovered a series of "leading indicators." In particular, she has identified a leading indicator for retail trade in furniture and appliances. Although the indicator is for Canada as a whole, Kate is wondering if it can give her some insight into Woodbon's sales.

Creating Graphs to Examine the Relationships Between the Response Variable and the Explanatory Variables

Kate begins by collecting data for the three additional (potential) explanatory variables. She finds Statistics Canada data for housing starts in New Brunswick[1] (Woodbon is located in Saint John, and delivers throughout the province). The data are available on a quarterly basis. Kate decides to add up the quarterly numbers so she can relate annual housing starts to annual sales.

Statistics Canada provides data on a variety of mortgage interest rates, and Kate decides to work with mortgage rates for five-year conventional mortgages at chartered banks. Statistics Canada provides data about monthly mortgage rates[2] (based on the last Wednesday of the month). Kate could simply use the mortgage rate for one month of the year to represent mortgage rates for that year (e.g., the January or June mortgage rates). In general, it is simplest to use the available data in raw form to build models. In this case, Kate decides to compute a simple average of the monthly rates to create a data series of annual average mortgage rates.

An excerpt of the data set, including data on sales and advertising, is shown on the next page in Exhibit 14.1. The complete data set is available in an Excel file called SEC14-2.

Initially, it can be useful to create scatter diagrams to explore the relationship between sales and each one of the potential explanatory variables. The four scatter diagrams are shown in Exhibit 14.2.

We have already established (in Chapter 13) that there is a positive association between Woodbon's advertising expenditure and sales. From the scatter diagrams we can see that there is a negative relationship between Woodbon sales and mortgage interest rates, as expected. There appears to be a positive relationship between Woodbon sales and the Canada-wide leading indicator for retail trade in furniture and appliances, although it may not be linear. There is a somewhat curved appearance to the plot, which flattens out for higher levels of the leading indicator.[3] There does not appear to be much of a relationship between Woodbon sales and housing starts in New Brunswick. However, this does not necessarily mean that housing starts will not be useful in the regression model. This variable, in conjunction with others, could still potentially improve the model's predictions.

---

[1] Statistics Canada, "CMHC, Housing Starts, under Construction and Completions, All Areas; New Brunswick; Housing Starts; Total Units; Unadjusted (units) [J15005]," CANSIM Table 027-0008, **www.statcan.gc.ca**, accessed October 13, 2008.

[2] Statistics Canada, "Financial Market Statistics, Last Wednesday Unless Otherwise Stated, Monthly (percent)(1), Bank of Canada – 7502 Rates (Percent) Chartered Bank – Conventional Mortgage: 5 year," CANSIM Table 176-0043, **www.statcan.gc.ca**, accessed October 13, 2008.

[3] It is possible to model non-linear relationships, but this is a more advanced topic.

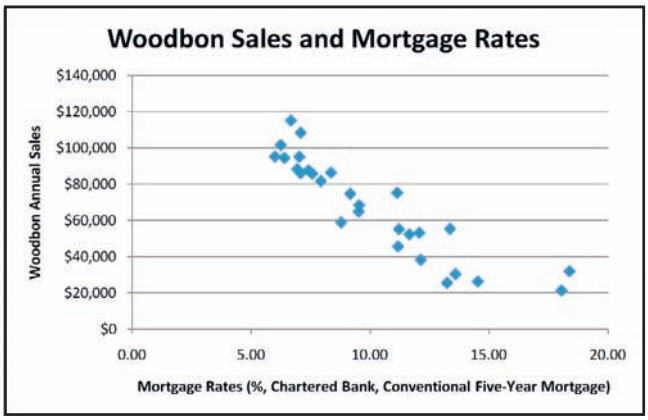
EXHIBIT 14.1

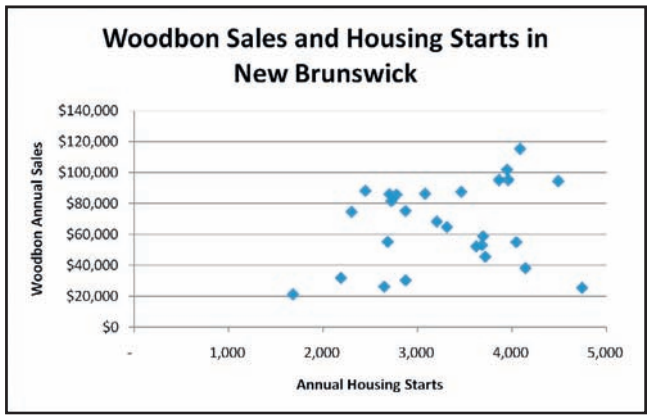
Woodbon Sales and Explanatory Variable Data

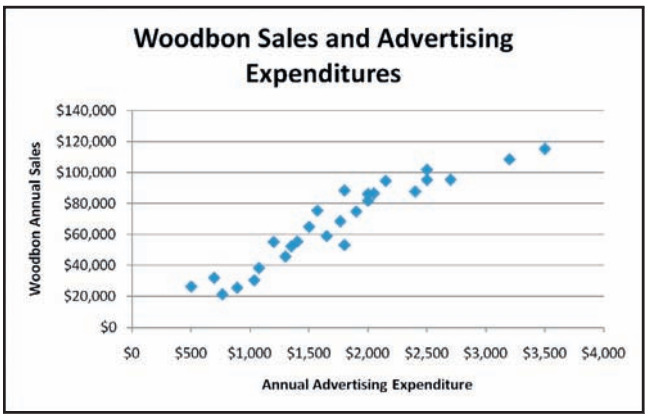
Woodbon Mortgage Rates, Housing Starts, Advertising Expenditure, Leading Indicator, and Sales, 1980–2007					
Year	Mortgage Rates	Housing Starts (New Brunswick)	Advertising Expenditure	Leading Indicator (Retail Trade, Furniture and Appliances, Canada)	Sales
1980	14.52083	2,646	$ 500	599	$ 26,345
1981	18.37500	2,188	$ 695	639	$ 31,987
1982	18.04167	1,680	$ 765	577	$ 21,334
⋮	⋮	⋮	⋮	⋮	⋮
2004	6.23333	3,947	$2,500	2,014	$101,760
2005	5.99167	3,959	$2,700	2,195	$ 95,400
2006	6.66250	4,085	$3,500	2,515	$115,320
2007	7.07083	4,242	$3,200	2,648	$108,550

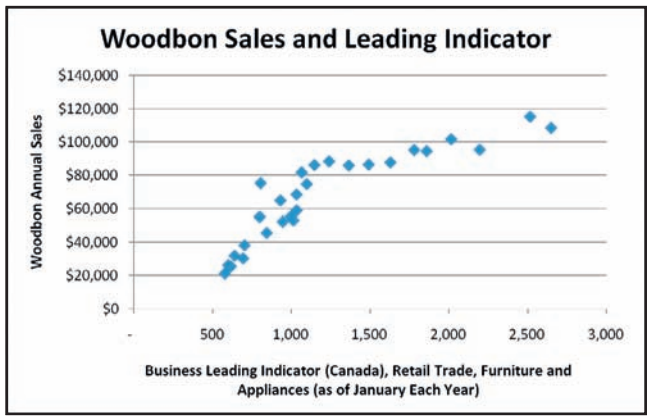
EXHIBIT 14.2

Scatter Diagrams for Woodbon Sales and Explanatory Variable Data







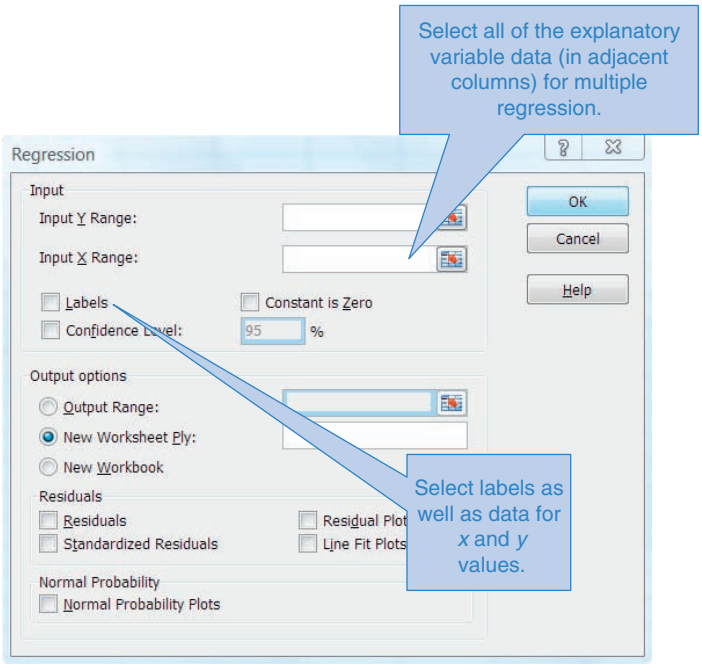


Determining the Relationship Between the Response Variable and the Explanatory Variables

We will begin our analysis by adding all of the new explanatory variables to create a new multiple regression model. The model is built using Excel's **Data Analysis Regression** tool, as illustrated in Chapter 13. The only difference is that more than one explanatory variable will be selected for **Input *X* Range:**. It is highly recommended that you include labels when selecting the data in Excel, because it will make the output much easier to read. Exhibit 14.3 illustrates.

EXHIBIT 14.3

Excel Regression Dialogue Box



The Woodbon output for the regression model is shown on the next page in Exhibit 14.4.

At first glance, this model looks promising. The R^2 value is 0.955. The mathematical relationship from the regression model is as follows:

$$\text{Woodbon annual sales} = \$89{,}159.92 - \$3{,}814.57\ (\text{mortgage rate}) - \$6.40\ (\text{housing starts}) + \$14.71\ (\text{advertising expenditure}) + \$10.44\ (\text{leading indicator})$$

How do we interpret this mathematical model?

1. The intercept \$89,159.92 is an estimate of average Woodbon sales when all of the explanatory variables have a value of zero. This number does not have a practical interpretation, because it is highly unlikely, for example, that mortgage rates would ever be zero. Additionally, a regression model should never be applied for values of the explanatory variables that are outside of the ranges in the data set used to build

EXHIBIT 14.4

Excel Regression Output for Woodbon Sales and Explanatory Variable Data (All Variables Included)

Regression Statistics	
Multiple R	0.977269651
R Square	0.955055971
Adjusted R Square	0.947239618
Standard Error	6251.003405
Observations	28

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	19097834678	4774458670	122.186906	3.83971E-15
Residual	23	898726002.1	39075043.57		
Total	27	19996560680			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	89159.92438	14444.22871	6.172702342	2.6863E-06	59279.7609	119040.0879
Mortgage Rates	-3814.56869	690.1923678	-5.52681958	1.275E-05	-5242.340378	-2386.79701
Housing Starts	-6.39732635	1.837777835	-3.48101181	0.00201934	-10.19905943	-2.59559327
Advertising Expenditure	14.71076331	6.518471372	2.256781148	0.03382727	1.226277974	28.19524865
Leading Indicator	10.44591571	6.900630411	1.513762524	0.14370781	-3.829125824	24.72095724

the model, and none of the variable values in the data set for Woodbon was even close to zero.

2. The mortgage rate coefficient can be interpreted as follows. If all values of the other variables are fixed at specific levels, there would be a decrease in Woodbon's average sales by $3,814.57 for each 1% increase in the conventional five-year mortgage rates at the chartered banks. It seems reasonable that higher mortgage rates would leave less money available to households for spending on furniture, and so the negative relationship makes sense.
3. The housing starts coefficient can be interpreted as follows. If all values of the other explanatory variables are fixed at specific levels, there would be a decrease in Woodbon's average sales by $6.40 for each additional housing start in New Brunswick. We would have expected that furniture spending would increase, not decrease, with additional housing starts. However, it is important to recognize that this coefficient applies only when all of the other explanatory variables are included in the model. There may be some interaction between housing starts and one or more of the other explanatory variables that results in a coefficient of the "wrong" sign. Remember, there did not appear to be a strong relationship between Woodbon's sales and housing starts in the first place. The fact that the sign on the coefficient is "wrong" increases our suspicion that this "explanatory" variable may not actually explain very much about Woodbon's sales.
4. The advertising expenditure coefficient indicates that each additional dollar in advertising spending results in an increase of $14.71 in Woodbon's annual sales, when all other variables held the same. Note that this coefficient is different from the $35.17 value in the regression model based on advertising expenditure alone (see Chapter 13 page 484).

It appears that the "all-in" model has some difficulties. Normally, we might stop and rethink at this point. However, we will continue analyzing this model, because it will give us the opportunity to discuss relationships which both do and do not meet the required conditions of the theoretical linear regression model.

DEVELOP YOUR SKILLS 14.1

MyStatLab

The Develop Your Skills exercises in this chapter frequently refer to a data set called "Salaries."

SALARIES

1. For the Salaries data set, create scatter diagrams showing the relationship between each possible explanatory variable and salaries. Are there any obvious problems? Are there some variables that seem particularly strong as candidates for explanatory variables?
2. For the Salaries data set, create a multiple regression model that includes all the possible explanatory variables. Interpret this model. Are there any obvious difficulties with this model?
3. Create a scatter diagram showing the relationship between age and years of experience. Does it seem sensible to include both of these explanatory variables in the model?
4. Create a multiple regression model for the Salaries data set that includes years of postsecondary education and age as explanatory variables. Interpret the model.
5. Create a multiple regression model for the Salaries data set that includes years of postsecondary education and years of experience as explanatory variables. Interpret the model.

14.2 CHECKING THE REQUIRED CONDITIONS

The Theoretical Model

In Chapter 13, we described how the least-squares line was created, as a best fit between the explanatory and response variables. The theoretical relationship was

$$y = \beta_0 + \beta_1 x + \varepsilon$$

This indicated that the y-value could be predicted from x. The ε term reminds us that we do not expect the prediction to be perfect. There may be some unexplained or random variation in the y-values that cannot be predicted from the x-values.

The corresponding notation for the regression relationship based on sample data is

$$\hat{y} = b_0 + b_1 x$$

The coefficients b_0 and b_1 were arrived at by minimizing the sum of the squared residuals for the data set, that is

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

Now we extend the model so that it includes more explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

The corresponding notation for the relationship based on sample data is

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

Again, the coefficients are estimated by minimizing the sum of squared residuals, for all of the data points in the data set. This requires the use of advanced algebra, but the idea is the same as in Chapter 13. Essentially, this creates a multiple regression model where the predicted values are simultaneously as close as possible to the observed values.

When the model has one response variable and one explanatory variable, as in Chapter 13, we can think of the relationship as a line, because we are operating in two dimensions (x and y). When we have one response variable and k explanatory variables, we are operating in $k+1$ dimensions. With two explanatory variables, we can imagine a plane as the regression surface. With more than two explanatory variables, there is no way to picture the regression relationship.

Examining the Residuals

In Chapter 13, we saw that analysis of the residuals ($y_i - \hat{y}_i$) was required to check whether the sample data appear to conform to the requirements of the least-squares regression model. As before, we can legitimately make predictions with the model, or perform hypothesis tests about the relationship between the y- and x-variables, only if these requirements are met.

Requirements for Predictions or Hypothesis Tests About the Multiple Regression Relationship

1. For any specific combination of the x-values, there are many possible values of y and the residual (or "error term") ε. The distribution of these ε-values must be normal for any specific combination of x-values. This means that the actual y-values will be normally distributed around the predicted y-values from the regression relationship, for every specific combination of x-values.
2. These normal distributions of ε-values must have a mean of zero. The actual y-values will have expected values, or means, that are equal to the predicted y-values from the regression relationship.
3. The standard deviation of the ε-values, which we refer to as σ_ε, is the same for every combination of x-values. The normal distributions of actual y-values around the predicted y-values from the regression relationship will have the same variability for every specific combination of x-values.
4. The ε-values for different combinations of the x-values are not related to each other. The value of the error term ε is statistically independent of any other value of ε.

As in Chapter 13, we create a number of residual plots to check these requirements. If they appear to be met in the sample data, we will assume they are met in the population. As usual, Excel is a great help in creating the required graphs. As before, in the **Regression** dialogue box, you should tick **Residuals**, **Standardized Residuals**, and **Residual Plots**. As in Chapter 13, you should create a histogram of the residuals, and you should plot the residuals against time if you have time-series data.

Variation in the Residuals Is Constant A plot of the residuals against the predicted values from the model can give us an indication of whether the variability of

the error term is constant. Such a plot can be created from the information created by Excel in the **Residual Output**, an excerpt of which is shown below in Exhibit 14.5 (note that some of the rows of data have been hidden in the worksheet).

EXHIBIT 14.5

Excerpt of Excel Regression Output for Woodbon, Showing Residuals

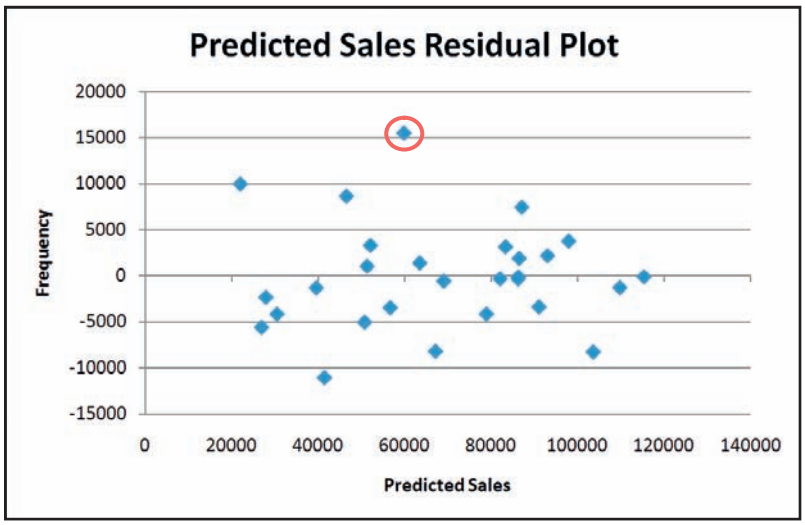
RESIDUAL OUTPUT

Observation	*Predicted Sales*	*Residuals*	*Standard Residuals*
1	30454.36781	-4109.37	-0.712267688
2	21968.79525	10018.2	1.736433405
3	26872.26659	-5538.27	-0.959935572
12	59828.247	15556.8	2.696417798
13	63476.35725	1449.64	0.251263391
25	97947.18162	3812.82	0.660867428
26	103625.1312	-8225.13	-1.425643909
27	115371.4319	-51.4319	-0.008914583
28	109785.5139	-1235.51	-0.214148915

The plot of residuals against predicted values is shown below in Exhibit 14.6. In Excel 2007, it is quite easy to produce this scatter diagram. Simply highlight the two adjacent columns (Predicted Sales and Residuals, in this case), and **Insert** a **Scatter** diagram.

EXHIBIT 14.6

Plot of Residuals Against Predicted Sales, Woodbon



The requirement is that the variability of the error term is constant, so a residual plot with constant variability (a horizontal band, centred on zero vertically) is ideal. This residual plot does not show any particular pattern, and appears as a horizontal

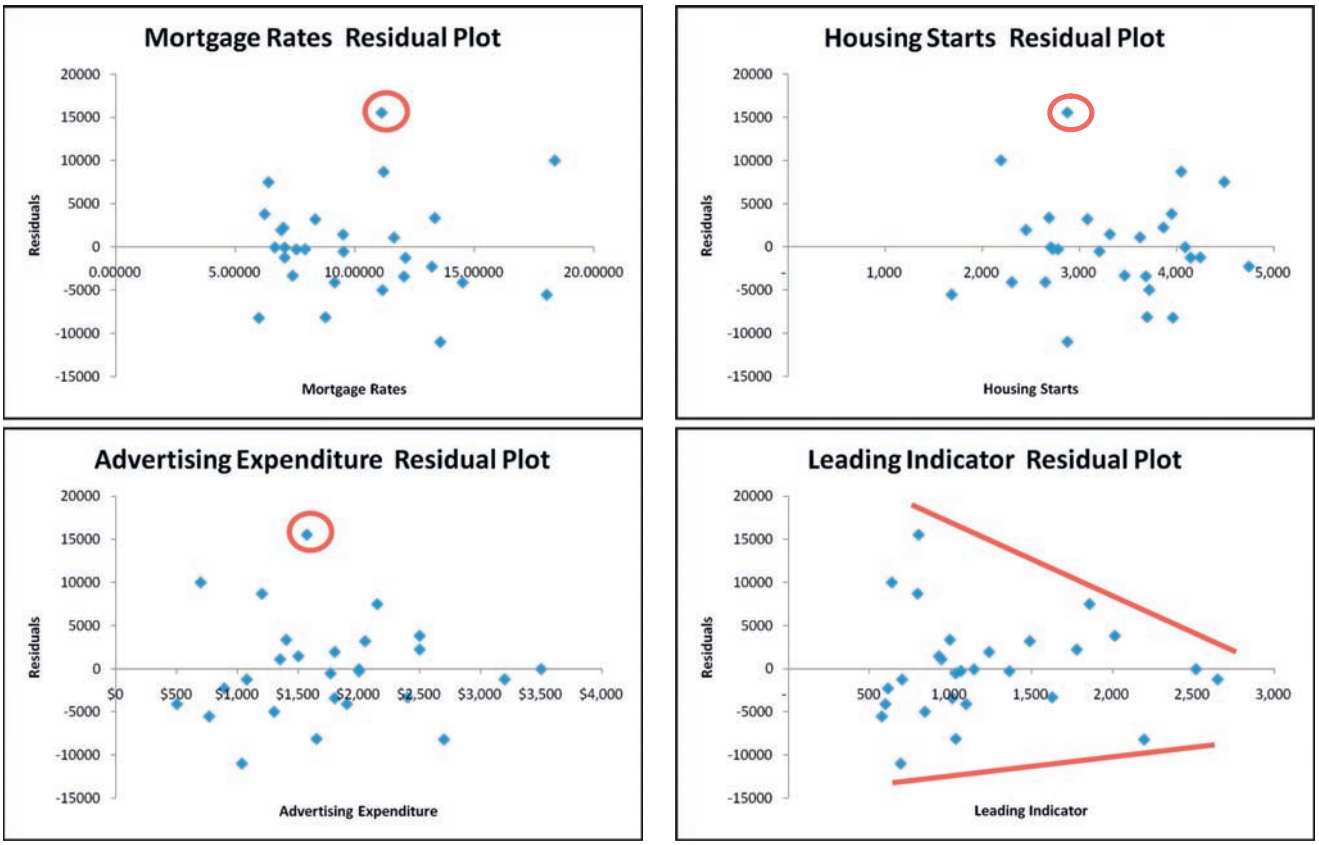
band. There is one residual that is unusually high (it is circled on the plot). This point corresponds to the data point for 1991.

It can also be helpful to plot the residuals against each individual *x*-variable, particularly if there appears to be a problem with the plot of the residuals against the predicted values. The additional plots can indicate which explanatory variables might be the source of any problem.

When **Residual Plots** is ticked as an option in Excel's **Data Analysis** tool for **Regression**, graphs are automatically created to show residuals against every *x*-variable in the model. The graphs for the Woodbon model are shown in Exhibit 14.7 below. Note that the graphs have been resized for visibility.

EXHIBIT 14.7

Residual Plots for Woodbon Multiple Regression Model



The mortgage rates residual plot has the desired horizontal band appearance, although there is one point (circled on the plot) where a residual seems unusually high. This is the data point for 1991, the same point that stood out in Exhibit 14.6.

The housing starts residual plot exhibits fairly constant variability in the residuals, although again there is one point (circled) that gives an unusually high residual. Again, this is the data point from 1991.

The advertising expenditure residual plot has something of a horizontal band appearance, although there does not seem to be as much variability in the residuals when the advertising expenditure is higher. Again, the data point for 1991 shows an unusually high residual.

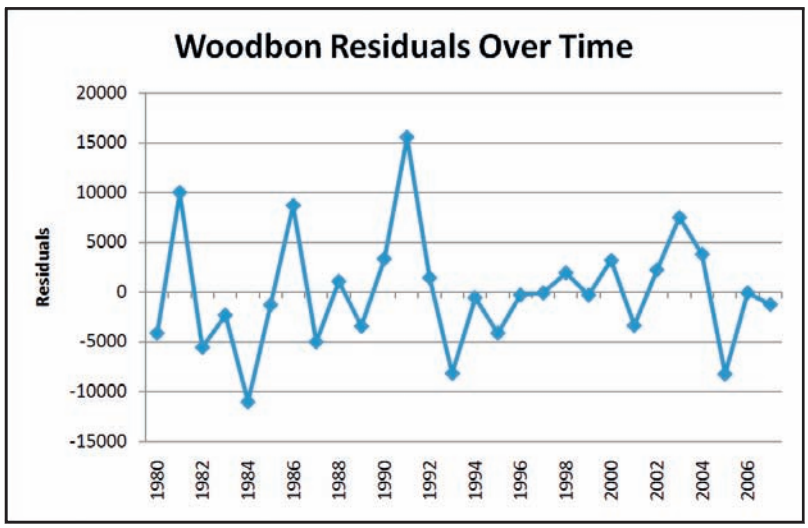
The leading indicator residual plot gives the greatest cause for concern. This plot shows reduced variability in the error term for higher values of the leading indicator, and the pattern is more pronounced than for the advertising expenditure residual plot, although it is also affected by the 1991 data point. Remember that the scatter plot of sales against the leading indicator looked non-linear. The residual plot is consistent with the curved scatter diagram that we saw earlier. Since we are trying to build a linear multiple regression model, we may not be able to use this variable in its present form.

Independence of Error Terms Plotting the residuals in the time order in which the data occurred allows us to check if the error terms are related over time. The Woodbon data set was arranged by year, so the residuals are also arranged by year.

A plot of the residuals over time is shown in Exhibit 14.8 below.

EXHIBIT 14.8

Plot of Residuals Over Time, Woodbon Model



There does not appear to be any particular pattern in the residuals over time, and so it is reasonable to conclude that the residuals are independent over time.

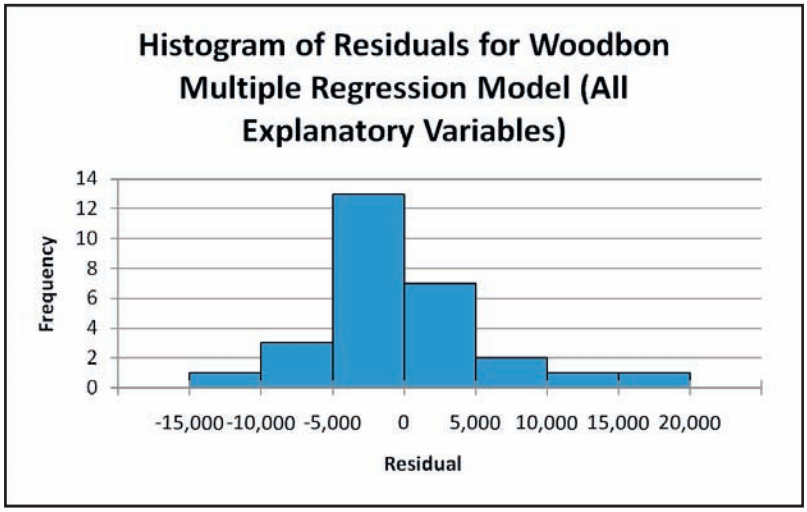
Normality of Residuals A histogram is created for the residuals, to check for normality. The histogram for the initial Woodbon model is shown on the next page in Exhibit 14.9.

The histogram appears to be approximately normal, although it is somewhat skewed to the right. As well, the histogram appears to be centred approximately around zero, which is desirable.

Outliers and Influential Observations Outliers, that is, observations that are far from other observations, should always be investigated. Such points may be the result of

EXHIBIT 14.9

Histogram of Residuals for Woodbon Multiple Regression Model



an error in observing or recording the data. If that is the case, and they are not corrected or removed, the result will be a model that is less correct than it could be. As before, a general rule is to investigate any observation with a standardized residual that is $\geq +2$ or ≤ -2. If you examine the **Residual Output** for the Woodbon model (see Exhibit 14.5 on page 533), you will see that there is one point that would be identified as an outlier, that is, observation 12. It is no surprise that data point 12 is the observation for 1991, given how often it has shown up as an unusual point in the residual plots. This data point is accurately recorded. There is no obvious reason why it does not belong in the data set. Therefore, we will not discard it.

An influential observation is one that has an extreme effect on the regression model. In the simple linear regression model discussed in Chapter 13, we could use scatter plots to identify such values. Influential observations are more difficult to locate in the multiple regression model, because the influence might come from just one x-variable, or a combination of them. There are several techniques available to help identify influential observations, and they can be found in more advanced texts. If you suspect that an observation is having an undue influence on the regression model, one way to check is to recalculate the model without the suspect observation. If the regression coefficients change significantly (a judgment call), then the observation is influential.

What If the Required Conditions Are Not Met? The hypothesis tests and confidence intervals that will be described in Sections 14.3 and 14.5 are valid only if the data appear to meet the requirements for the linear regression model. If they do not, further work must be done before hypothesis tests are done or confidence intervals are calculated.

There exist more advanced techniques that could be used to solve some of the problems that arise. For example, it may be possible to transform the data by applying some mathematical function (such as a logarithm or square root) to the original data, and work with the new measurements. It may also be possible to build a useful model without including the variables that are responsible for the conditions not being met. It may be necessary

to start over, to try to find explanatory variables that meet the requirements. It may also be useful to proceed, if the violation of the required conditions is not too pronounced. If the resulting model provides useful predictions, it may be the best we can do.

Before we proceed with our regression analysis, we will remove the leading indicator as an explanatory variable, as it does not meet the requirements for linear regression in its present form. In particular, the variability in the residuals is not constant. Especially when we are just beginning our analysis of a model, we should not necessarily discard explanatory variables that do not meet the requirements of a linear regression model, especially when they seem to be reasonable choices. As mentioned above, we may be able to transform the leading indicator data so that the model does meet the requirements for linear regression. However, the leading indicator data also presents other difficulties (see Section 14.5), so we will drop it now to streamline the discussion. As we will see later, choosing the best explanatory variables for any model is an art.

Once we use Excel to create a new regression model, we will see that the model better meets the required conditions for linear regression. Example 14.2 illustrates.

EXAMPLE 14.2

Checking conditions for linear multiple regression

Use Excel to re-specify the multiple regression relationship between Woodbon sales and mortgage rates, housing starts, and advertising expenditure. Check to see that the new model meets the required conditions for hypothesis tests and confidence intervals.

The regression output for the new model is shown in Exhibit 14.10 below.

EXHIBIT 14.10

Regression Output for Woodbon Model, with Mortgage Rates, Housing Starts, and Advertising Expenditure as Explanatory Variables

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.974976011
R Square	0.950578223
Adjusted R Square	0.944400501
Standard Error	6416.987757
Observations	28

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	19008295115	6336098372	153.871961	8.35857E-16
Residual	24	988265564.9	41177731.9		
Total	27	19996560680			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	80640.56739	13655.93989	5.9051642	4.3041E-06	52456.09289	108825.0419	52456.09289	108825.0419
Mortgage Rates	-3521.91472	680.1559808	-5.1780986	2.649E-05	-4925.68766	-2118.14178	-4925.68766	-2118.14178
Housing Starts	-5.47726	1.780411941	-3.0763977	0.00517198	-9.151844821	-1.80266558	-9.15184482	-1.80266558
Advertising Expenditure	23.41351	3.153796991	7.42391066	1.1541E-07	16.90439008	29.92262413	16.90439008	29.92262413

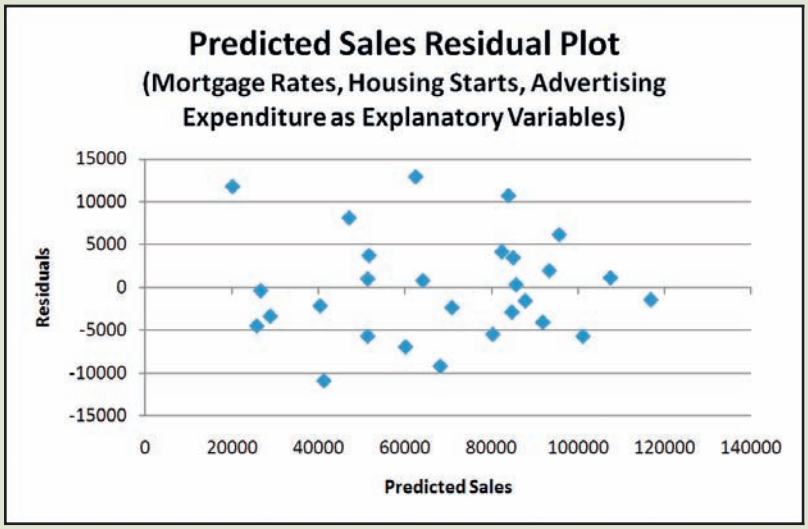
From the output, we can see that the regression relationship has become (approximately):

$$\begin{aligned}\text{Woodbon annual sales} = \$80{,}640.57 &- \$3{,}521.91\ (\text{mortgage rate}) \\ &- \$5.48\ (\text{housing starts}) + \$23.41\ (\text{advertising expenditure})\end{aligned}$$

The various residual plots for the revised model all appear to conform to the required conditions. Exhibit 14.11 shows the revised predicted sales residual plot, which appears to have the desirable horizontal band of points.

EXHIBIT 14.11

Predicted Sales Residual Plot for Woodbon Model, with Mortgage Rates, Housing Starts, and Advertising Expenditure as Explanatory Variables



As well, none of the residual plots for the three explanatory variables give an indication of violation of the required conditions. Exhibit 14.12 opposite shows these residual plots.

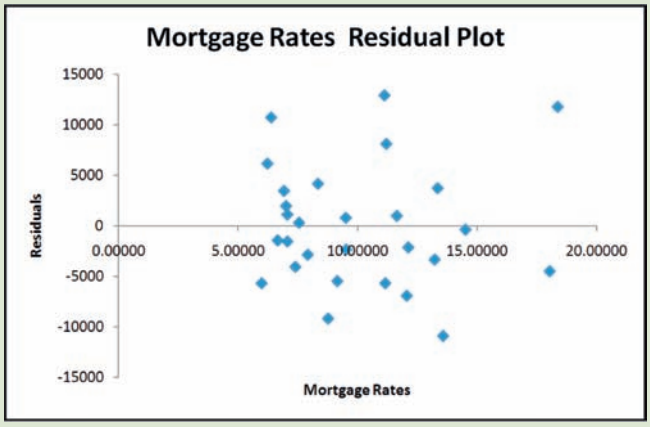
The plot of the residuals over time does not exhibit any particular pattern, so we can conclude that the residuals are independent over time. Exhibit 14.13 illustrates.

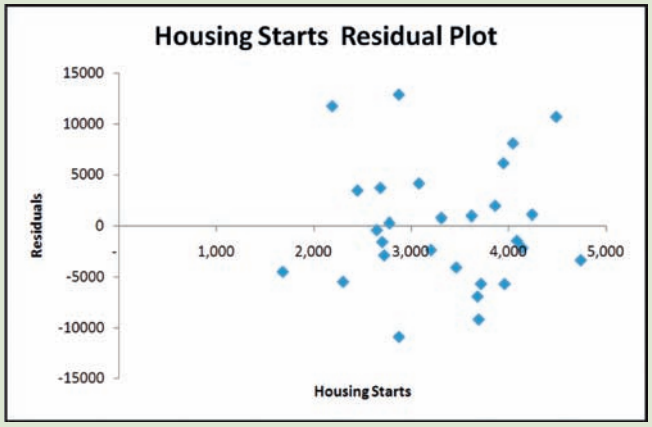
Finally, the histogram of residuals for the new Woodbon model appears fairly normal, although there is some right-skewness. Exhibit 14.14 on page 540 illustrates.

A check of the standardized residuals reveals one data point that could be classified as an outlier. This is the data point that corresponds to 1991 (notice that this point also attracted our attention when we examined the residual plots for the original model). Since the data are correct, we will leave the point in the data set. It appears that 1991 was not a typical year for Woodbon.

EXHIBIT 14.12

Residual Plots for Woodbon Multiple Regression Model (Mortgage Rates, Housing Starts, Advertising Expenditure as Explanatory Variables)





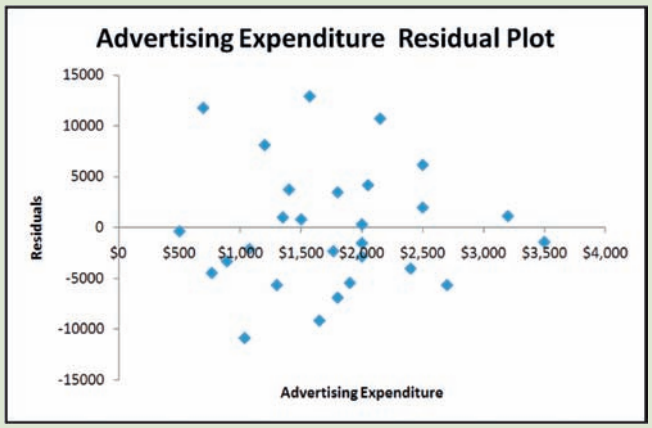


EXHIBIT 14.13

Plot of Residuals Over Time, Woodbon Model (Mortgage Rates, Housing Starts, Advertising Expenditure as Explanatory Variables)

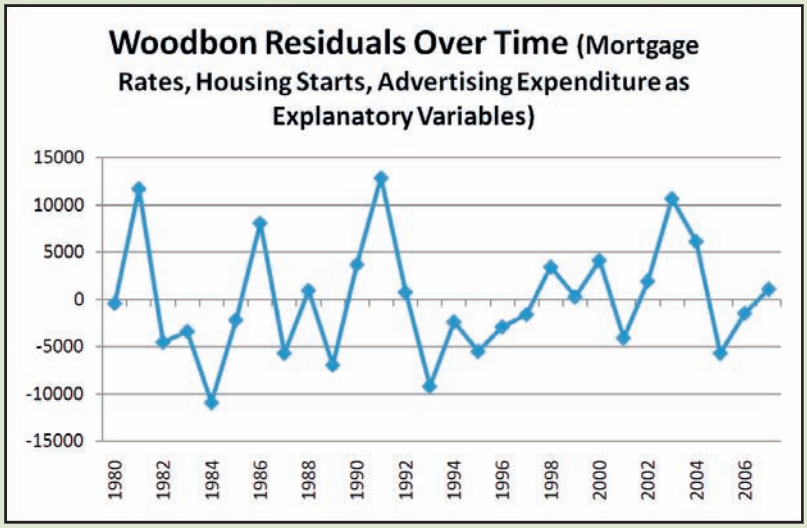
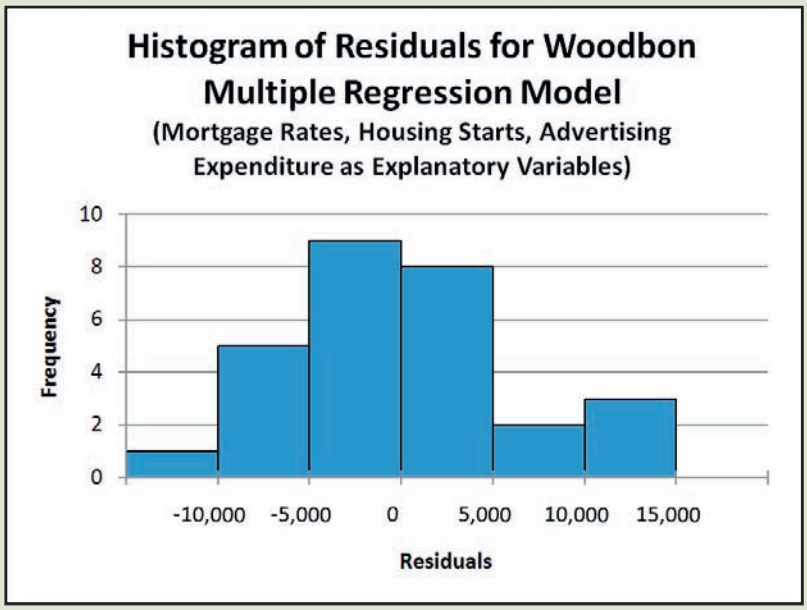


EXHIBIT 14.14

Histogram of Residuals for Woodbon Model (Mortgage Rates, Housing Starts, Advertising Expenditure as Explanatory Variables)



The new Woodbon model appears to meet all of the required conditions for the linear regression model. However, it still does not meet the test of common sense, in that the coefficient for housing starts is negative when we would expect it to be positive. We will say more about this difficulty in Section 14.5, when we discuss criteria for selecting explanatory variables.

GUIDE TO TECHNIQUE

Checking Requirements for the Linear Multiple Regression Model

When:

- before performing hypothesis tests or using the regression relationship to create confidence or prediction intervals
- using sample data to assess whether the relationship conforms to requirements

Steps:

1. Produce scatter diagrams for the relationship between each explanatory variable and the response variable. Check to see that each relationship appears linear (no pronounced curvature).
2. Use Excel's **Regression** tool to produce the **Residuals**, **Standardized Residuals**, and **Residual Plots**.

3. Create a plot of the residuals versus the predicted *y*-values. The residuals should be randomly distributed around zero, with the same variability throughout.
4. Examine the plots of residuals versus each explanatory variable. Again, the residuals should be randomly distributed around zero, with the same variability (a random horizontal band appearance is desirable).
5. Check time-series data by plotting the residuals in time order. There should be no discernible pattern to the plot.
6. Create a histogram of the residuals. This should be approximately normal, and centred on zero.
7. Check for outliers and influential observations. Carefully check any data point with a standardized residual $\geq +2$ or ≤ -2.

Note: If these investigations indicate significant problems, you should not proceed with a hypothesis test of the significance of the model, and you should not create confidence intervals or prediction intervals with the model in its current form.

DEVELOP YOUR SKILLS 14.2

MyStatLab

The Develop Your Skills exercises in this chapter frequently refer to a data set called "Salaries."

SALARIES

6. Examine the residual plots produced by Excel for the Salaries multiple regression model that you built for Develop Your Skills 14.1, Exercise 2, which included all possible explanatory variables. Are these residual plots consistent with the required conditions? Create a plot of the residuals versus predicted salaries. Does this plot meet the required conditions?
7. Examine the residual plots produced by Excel for the Salaries multiple regression model that you built for Develop Your Skills 14.2, Exercise 4, which included years of postsecondary education and age as explanatory variables. Are these residual plots consistent with the required conditions? Create a plot of the residuals versus predicted salaries. Does this plot meet the required conditions?
8. Examine the residual plots produced by Excel for the Salaries multiple regression model that you built for Develop Your Skills 14.2, Exercise 5, which included years of postsecondary education and years of experience as explanatory variables. Are these residual plots consistent with the required conditions? Create a plot of the residuals versus predicted salaries. Does this plot meet the required conditions?
9. Create histograms of the residuals for the models discussed in Exercises 6, 7, and 8 above. Do these histograms appear to be at least approximately normal?
10. Check the Excel output for the models created in Exercises 6, 7, and 8 above for outliers. If you had access to the original records for this data set, what would you do?

14.3 HOW GOOD IS THE REGRESSION?

Because the new Woodbon model appears to meet the required conditions, we can now conduct hypothesis tests about the overall model, and about the individual explanatory variables. We begin by testing whether the regression model is significant. Given this sample data set, is there evidence that there is a population regression relationship between sales and at least one of the explanatory variables?

Is the Regression Model Significant?—The F-Test

In the discussion of simple linear regression (Chapter 13), we performed a test of hypothesis about the slope of the regression line.

In multiple regression, we test the model as a whole.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1: At least one of the β_i's is not zero.

If the null hypothesis is true, then the y-variable is not related to any of the x-variables. If the alternative hypothesis is true, then the y-variable is related to at least one of the x-variables. As in Chapter 13, we hope to reject the null hypothesis, so that we can conclude there is a significant relationship between the response variable and at least one of the explanatory variables.

We conduct the hypothesis test by examining how much of the variation in the y-variable is explained by the regression relationship. Remember from Chapter 13 that the total variation in the y-values can be broken down into two parts: the variation that is explained by the regression relationship, and the variation that is left unexplained.

$$\Sigma(y - \bar{y})^2 = \Sigma(\hat{y} - \bar{y})^2 + \Sigma(y - \hat{y})^2$$

It is usual to describe this relationship as follows:

$$SST = SSR + SSE$$

The total sum of squares (SST) is equal to the sum of squares explained by the regression (SSR) plus the residual (or error) sum of squares (SSE).

When the response variable (y) is related to the explanatory variables, then SSR will be relatively large, and SSE will be relatively small. Before we can compare SSR and SSE, we must adjust them so they are directly comparable. This is accomplished by dividing each by its degrees of freedom to calculate the associated mean square value.[4]

The degrees of freedom for the error sum of squares are $n - (k + 1)$, because we estimate k coefficients plus an intercept from n data points. The degrees of freedom for the total variation are $(n - 1)$. This leaves k degrees of freedom for the regression sum of squares.

The test statistic is the ratio of the mean squares, and is an F statistic. The F distribution is described on pages 412–419 in Chapter 11.

$$F = \frac{\dfrac{SSR}{k}}{\dfrac{SSE}{n - (k + 1)}} = \frac{MSR}{MSE}$$

[4] If this seems familiar, it should. We made the same adjustment to compare mean squares in Chapter 11. See page 411.

When the null hypothesis is true, and the response variable is not related to any of the explanatory variables, the mean square for the regression (MSR) will not be significantly larger than the mean square for error (MSE), and the F statistic will be relatively small. However, when the response variable is related to at least one of the explanatory variables, the MSR will be significantly larger than the MSE, and the F statistic will be relatively large. The question is, how large does the F statistic have to be to provide evidence of a significant relationship?

Of course, the answer depends on sampling variability. As usual, we have to refer to the sampling distribution of the F statistic to decide whether any specific F statistic is unusual enough for us to reject the null hypothesis. The sampling distribution of the F statistic depends on the number of data points and the number of explanatory variables.

The Sampling Distribution of $\frac{MSR}{MSE}$ in Linear Multiple Regression Models

The sampling distribution of $\frac{MSR}{MSE}$ follows the F distribution, with $(k, n - (k + 1))$ degrees of freedom, where n is the number of observed data points and k is the number of explanatory variables in the model.

Fortunately, the Excel output not only calculates the F statistic for the hypothesis test of the regression model, it also calculates the associated p-value.

EXAMPLE 14.3A

Hypothesis test of significance of regression model

Complete the hypothesis test for the significance of the Woodbon model based on mortgage rates, housing starts, and advertising expenditure. Use a 5% level of significance.

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1: At least one of the β_i's is not zero.

$\alpha = 0.05$

The Excel output of the regression model is reproduced on the next page in Exhibit 14.15, for ease of reference.

From the output, we see that F = 153.9 and the p-value is 0.000000000000000836. Since the p-value is less than the level of significance, we reject H_0. There is strong evidence to infer that there is a significant relationship between Woodbon sales and at least one of the explanatory variables. As always, we remember that a significant relationship is not necessarily a cause-and-effect relationship. Some other factor may be the cause of associated changes in sales and the explanatory variables.

EXHIBIT 14.15

Excel Regression Output for Woodbon Model, with Mortgage Rates, Housing Starts, and Advertising Expenditure as Explanatory Variables

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.974976011				
R Square	0.950578223				
Adjusted R Square	0.944400501				
Standard Error	6416.987757				
Observations	28				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	19008295115	6336098372	153.8719615	8.35857E-16
Residual	24	988265564.9	41177731.87		
Total	27	19996560680			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	80640.56739	13655.93989	5.905164203	4.30411E-06	52456.09289
Mortgage Rates	-3521.91472	680.1559808	-5.178098581	2.64897E-05	-4925.68766
Housing Starts	-5.47726	1.780411941	-3.076397702	0.005171979	-9.151844821
Advertising Expenditure	23.41351	3.153796991	7.423910661	1.15414E-07	16.90439008

Are the Explanatory Variables Significant?—The *t*-Test

If the hypothesis test of the overall regression model indicates a significant relationship between the response variable and at least one of the explanatory variables, the next step is to figure out which of the explanatory variables is significant.

We conducted a *t*-test about the slope of the regression line in Chapter 13. We test the individual coefficients in the multiple regression model in a similar fashion. The test of the coefficient of explanatory variable *i* is conducted as follows.

$$H_0: \beta_i = 0$$

$$H_1: \beta_i \neq 0$$

The test statistic is $t = \dfrac{b_i}{s_{b_i}}$, with $(n - (k + 1))$ degrees of freedom.

It is important to realize that this *t*-test for the significance of each explanatory variable assumes that all the other explanatory variables are included in the model. The *p*-values for the individual coefficients do give us some indication of how important each explanatory variable is. Those with small *p*-values are likely more strongly related to the response variable. However, we cannot just eliminate an explanatory variable

with a *p*-value greater than the level of significance. If we decide to eliminate any explanatory variable, we must rerun the regression analysis and examine the new model and the new *p*-values for the coefficients.

Remember that the *t*-tests for the individual coefficients should only be conducted if the F-test of the overall model shows that it is significant. We can control the Type I error rate on a single *t*-test with the level of significance (α), but the error rate becomes larger with repeated tests based on the same data set.[5] Therefore, the individual *t*-tests should only be performed when the overall Type I error rate is controlled, through the F-test.

EXAMPLE 14.3B

Hypothesis tests of individual coefficients in regression model

Conduct hypothesis tests about the significance of the individual coefficients in the Woodbon model.

As in the simple linear regression case, the Excel output contains the *p*-values for the two-tailed tests of significance for the individual coefficients. An excerpt from the Excel output is shown below in Exhibit 14.16.

EXHIBIT 14.16

Excerpt from the Excel Regression Output for Woodbon Model, with Mortgage Rates, Housing Starts, and Advertising Expenditure as Explanatory Variables

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	80640.56739	13655.93989	5.9051642	4.3041E-06
Mortgage Rates	-3521.91472	680.1559808	-5.1780986	2.649E-05
Housing Starts	-5.47726	1.780411941	-3.0763977	0.00517198
Advertising Expenditure	23.41351	3.153796991	7.42391066	1.1541E-07

For convenience, we will refer to the mortgage rates as explanatory variable 1, housing starts as explanatory variable 2, and advertising expenditure as explanatory variable 3.

Hypothesis test for mortgage rates:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$t = \frac{b_1}{s_{b_1}} = -5.18 \text{ (from Excel output)}$$

The *p*-value is 0.000026, so we reject H_0. There is strong evidence that mortgage rates are a significant explanatory variable for Woodbon annual sales, when housing starts and advertising expenditure are included in the model.

[5] This problem was discussed in Chapters 10 and 11—be careful when you are skating back and forth across the frozen lake!

Hypothesis test for housing starts:

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

$t = \dfrac{b_2}{s_{b_2}} = -3.08$ (from Excel output)

The *p*-value is 0.005, so we reject H_0. There is strong evidence that housing starts are a significant explanatory variable for Woodbon annual sales, when mortgage rates and advertising expenditure are included in the model.

Hypothesis test for advertising expenditure:

$H_0: \beta_3 = 0$

$H_1: \beta_3 \neq 0$

$t = \dfrac{b_3}{s_{b_3}} = 7.42$ (from Excel output)

The *p*-value is 0.0000001, so we reject H_0. There is strong evidence that advertising expenditure is a significant explanatory variable for Woodbon annual sales, when mortgage rates and housing starts are included in the model.

As always, while we can conclude that mortgage rates, housing starts, and advertising expenditure are significant explanatory variables for Woodbon annual sales, this does not mean that we can conclude that changes in these variables have caused the changes in sales.

Adjusted Multiple Coefficient of Determination

In Chapter 13, we used the coefficient of determination, or R^2, as an indication of how well the *x*-variable explained the variations in the *y*-variable. Remember, that

$$R^2 = \frac{SSR}{SST}$$

Adding more explanatory variables to the regression model will never reduce the R^2 value, and generally will tend to increase it. In fact, if you have *n* data points, it is always possible to develop a model that will fit the data perfectly, with $n - 1$ explanatory variables. However, this model is not likely to yield good predictions, because it is only the result of a lot of arithmetic instead of good thinking about the relationships between the response and explanatory variables. Such a model is usually described as "overfitted." It is possible to adjust the R^2 value to compensate for this tendency of R^2 to increase when another explanatory variable is added to the model.

It is easiest to see the relationship between the R^2 and the adjusted R^2 if we start with a restatement of R^2. We know SST = SSR + SSE, so SSR = SST – SSE. Substituting this into the formula for R^2 yields the following:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

The adjusted R^2 is calculated as follows:

$$\text{Adjusted } R^2 = 1 - \frac{\frac{\text{SSE}}{n-(k+1)}}{\frac{\text{SST}}{n-1}}$$

The adjusted R^2 is calculated by Excel as part of the **Regression** output.

The adjusted R^2 value will generally be smaller than the unadjusted R^2 value. As well, because the formula takes into account the number of explanatory variables being used (k), the adjusted R^2 will not necessarily increase when another variable is added to the model.

Excel's **Regression** output provides the adjusted R^2 value. For the Woodbon model, it is 0.944, as shown below in Exhibit 14.17. Notice that the adjusted R^2 value, at 0.944, is less than the R^2 value of 0.951.

EXHIBIT 14.17

Excel Regression Statistics for Woodbon Model, with Mortgage Rates, Housing Starts, and Advertising Expenditure as Explanatory Variables

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0.974976011
R Square	0.950578223
Adjusted R Square	0.944400501
Standard Error	6416.987757
Observations	28

DEVELOP YOUR SKILLS 14.3

The Develop Your Skills exercises in this chapter frequently refer to a data set called "Salaries."

11. Apply the formula for the adjusted R^2 to verify the value shown in Exhibit 14.17. Note that the SSE and SST are shown in the Excel **Regression** output.

SALARIES

12. Conduct a test of the significance of the overall model for the salaries model which includes all explanatory variables. Test the significance of the individual explanatory variables. What does this tell you?

13. Conduct a test of the significance of the overall model for the salaries model that includes years of postsecondary education and age as explanatory variables. Test the significance of the individual explanatory variables. What does this tell you?

14. Conduct a test of the significance of the overall model for the salaries model that includes years of postsecondary education and years of experience as explanatory variables. Test the significance of the individual explanatory variables. What does this tell you?

15. Compare the adjusted R^2 values for the three models of salary from Exercises 12, 13, and 14 above. Based on the adjusted R^2, which model does not seem worth considering at this point?

14.4 MAKING PREDICTIONS

One of the reasons for building a multiple regression model for Woodbon annual sales was to allow Kate Cameron, Woodbon's owner, to make sales predictions. Of course, the only explanatory variable in the present model that Kate can control is advertising expenditure. Kate will have to guess at the values of the other variables (mortgage rates and housing starts) if she wants to predict sales for the coming year. Suppose Kate plans to spend $3,000 on advertising next year, and she expects that five-year mortgage rates will be around 7% and that housing starts for the province will be 3,800.

By substituting these specific values into the regression equation, Kate arrives at a point estimate for Woodbon annual sales.

$$\begin{aligned}\text{Woodbon annual sales} &= \$80{,}640.56739 - \$3{,}521.91472\ (\text{mortgage rate}) - \$5.47726\ (\text{housing starts}) + \$23.41351(\text{advertising expenditure})\\ &= \$80{,}640.56739 - \$3{,}521.91472\ (7) - \$5.47726\ (3{,}800) + \$23.41351\ (3{,}000)\\ &= \$105{,}414.11\end{aligned}$$

In Chapter 13, we also created prediction and confidence intervals from regression relationships. Remember, a regression prediction interval predicts a particular value of y (sales) for a set of specific values of the x-variables (in this case, mortgage rate, housing starts, and advertising expenditure). A regression confidence interval predicts the average y for a set of specific values of the x-variables.

While the formulas for prediction and confidence intervals were fairly simple to understand when there was only one explanatory variable (with a given value of x_0), they become more complicated with two or more explanatory variables. Constructing these intervals for multiple regression requires the use of matrix algebra. An Excel add-in (**Multiple Regression Tools**) has been created to do these calculations. This add-in was first introduced in Chapter 13 (see page 514).

You should type the specific values of the explanatory variables that will be the basis of your intervals into adjacent columns in the spreadsheet containing the sample data. It is easiest to input the values in the correct order if they are typed at the bottom of the columns of explanatory variable data. Exhibit 14.18 illustrates (note that some of the rows of data have been hidden in the worksheet).

EXHIBIT 14.18

Typing in Specific Values of Explanatory Variables as Basis for Intervals

	A	B	C	D	E	F	G	H
1	Woodbon							
2	Year	Mortgage Rates	Housing Starts	Advertising Expenditure	Sales			
3	1980	14.52083	2,646	$500	$26,345			
4	1981	18.37500	2,188	$695	$31,987			
5	1982	18.04167	1,680	$765	$21,334			
6	1983	13.22917	4,742	$890	$25,584			
27	2004	6.23333	3,947	$2,500	$101,760			
28	2005	5.99167	3,959	$2,700	$95,400			
29	2006	6.66250	4,085	$3,500	$115,320			
30	2007	7.07083	4,242	$3,200	$108,550			
31		7.00000	3,800	$3,000	Specific Values of Explanatory Variables for Intervals			

As before, the **Confidence Interval and Prediction Intervals – Calculations** tool in **Multiple Regression Tools** requires you to indicate the locations of the labels and values of the variables, the location of the specific values of the explanatory variables on which you want to base your intervals, a level of confidence (percentage form), and an output range. Example 14.4 below provides the results.

EXAMPLE 14.4

Calculating confidence and prediction intervals with Excel

Use Excel to create a prediction interval for Woodbon sales, when mortgage rates are 7%, housing starts are 3,800, and advertising expenditure is $3,000.

Once these specific values are typed into the spreadsheet, the **Multiple Regression Tools** add-in is used to create the following output (note that columns have been resized for visibility).

EXHIBIT 14.19

Confidence Interval and Prediction Intervals – Calculations Result for Woodbon Data

Confidence Interval and Prediction Intervals - Calculations							
Point	95	= Confidence Level(%)		Prediction Interval		Confidence Interval	
Number	Mortgage Rates	Housing Starts	Advertising Expenditure	Lower limit	Upper limit	Lower limit	Upper limit
1	7	3800	3000	91016.42	119811.8113	99767.012	111061.2202

A 95% prediction interval for Woodbon sales when mortgage rates are 7%, housing starts are 3,800, and advertising expenditure is $3,000, is ($91,016.42, $119,811.81). Notice that even though we have added explanatory variables to the model, and the fit is better than it was in the simple linear regression model, the prediction interval is still quite wide.

Remember, it is not legitimate to make predictions from the regression model for values of the explanatory variables that are outside the range of the values in the data set on which the model is based. So, for example, Kate should not rely on the model to make predictions for an advertising budget of $5,000, because Woodbon has never spent more than $3,500 on advertising in the past.

DEVELOP YOUR SKILLS 14.4

MyStatLab

The Develop Your Skills exercises in this chapter frequently refer to a data set called "Salaries."

16. Use Excel to create a 95% confidence interval of average Woodbon sales, when mortgage rates are 6%, housing starts are 3,500, and advertising expenditure is $3,500. Interpret the interval.

17. Would it be appropriate to use the Woodbon model to make a prediction for mortgage rates of 6%, housing starts of 2,500, and advertising expenditure of $4,000? Explain why or why not.

18. Use the salaries model based on years of postsecondary education and age to make a 95% prediction interval estimate of the salary of an individual who is 35 years old and has five years of postsecondary education.

SALARIES

19. Use the salaries model based on years of postsecondary education and age to make a 95% confidence interval estimate of the average salary for individuals who are 35 years old and have five years of postsecondary education. Do you expect this confidence interval to be wider or narrower than the prediction interval estimate from Exercise 18? Why?

20. Use the salaries model based on years of postsecondary education and years of experience to make a 95% prediction interval estimate of the salary of an individual who has five years of postsecondary education and 10 years of experience.

14.5 SELECTING THE APPROPRIATE EXPLANATORY VARIABLES

Let us recap the steps we have followed to build the Woodbon sales model.

1. We began by thinking carefully about what explanatory variables might reasonably be expected to have an impact on Woodbon's sales, and we examined the relationship between each of these variables and sales.

2. We used Excel to estimate the multiple regression relationship between Woodbon sales and mortgage rates, housing starts, and a leading indicator for retail trade in furniture and appliances. We noted that the initial model had a negative coefficient for housing starts, the opposite of what we expected.
3. We used Excel to check whether the regression model met the required conditions. We examined plots of residuals against predicted sales, and residuals against each of the explanatory variables. We noted that the residual plot for the leading indicator did not exhibit the desired horizontal band shape. We also examined a plot of residuals over time, which did not exhibit any particular time-related pattern. We created a histogram of the residuals, and it appeared normal. We identified one outlier in the data set.
4. The model that included the leading indicator did not conform to the required conditions for linear regression (the variability in the residuals was not constant). For this and other reasons, we eliminated the leading indicator as an explanatory variable and recalculated the model. Again, we checked the required conditions, and this time did not identify any obvious violations of the required conditions. We identified one potential outlier, but because the associated data point was correct, we left it in the model.
5. We conducted an F-test of the significance of the overall model, and we were able to conclude there was a significant relationship between Woodbon sales and at least one of the remaining explanatory variables (mortgage rates, housing starts, and advertising expenditure).
6. We then conducted hypothesis tests about the coefficients for each explanatory variable. In each case, we concluded that there was evidence that each of the explanatory variables was significant, assuming the other explanatory variables were included in the model.
7. We examined the adjusted R^2 value, which was 0.94, indicating that the regression model explained a significant portion of the variability in Woodbon sales.

At this point, it might seem that we have done enough work and that we have the best possible model for Woodbon's sales, based on these explanatory variables. However, we must reflect carefully before we settle for the new model. More complicated models are not necessarily better. We should realize that the adjusted R^2 value for the new model, at 0.94, is not that much higher than the adjusted R^2 of 0.88 for the model based on advertising expenditure alone. This might not be enough of an improvement to justify the extra data required.

As well, it is possible for the adjusted R^2 value to be high for a model that does not predict well. Ultimately, a model that does not provide useful predictions of future sales for Woodbon will not be worth maintaining.

Building a good regression model is a process that requires many steps, as we have seen. Fortunately, computers do the calculations easily and quickly, and so we can look at a number of possible models before choosing the "best" one. It is not always easy to determine which is the best regression model, but these are some goals to keep in mind.

Goals for Regression Models

1. The model should be easy to use. It should be reasonably easy to acquire data for the model's explanatory variables.
2. The model should be reasonable. The coefficients should represent a reasonable cause-and-effect relationship between the response variable and the explanatory variables.
3. The model should make useful and reliable predictions. Prediction and confidence intervals should be reasonably narrow.
4. The model should be stable. It should not be significantly affected by small changes in explanatory variable data.

One method of finding the "best" possible regression model is to create regressions for all possible combinations of the explanatory variables being considered. For the Woodbon data set, this would entail creating a regression model for each of the following:

1. sales and mortgage rates
2. sales and housing starts
3. sales and advertising expenditure
4. sales and housing starts and mortgage rates
5. sales and mortgage rates and advertising expenditure
6. sales and housing starts and advertising expenditure
7. sales and housing starts and mortgage rates and advertising expenditure.

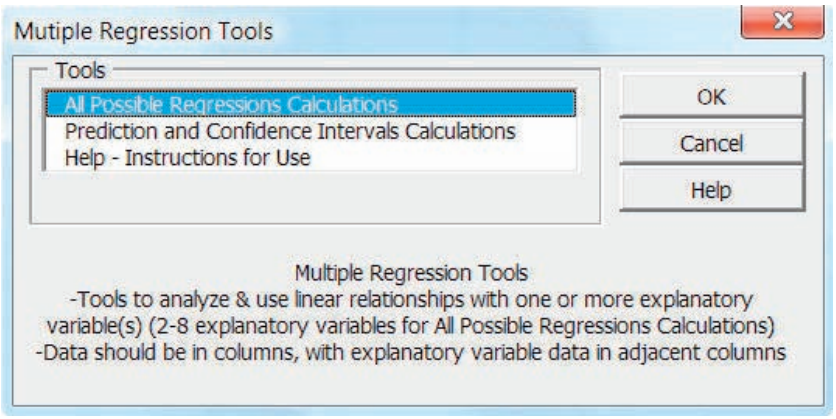
The resulting models can be assessed according to the criteria above. Some values that may be useful to compare models are as follows:

1. The adjusted R^2 values provide a measure of the strength of the relationship between the explanatory variables and the response variable.
2. The standard error (s_ε) gives some indication of how wide the confidence and prediction intervals would be. As discussed in Chapter 13, s_ε is the sample estimate of the standard deviation of the error terms (residuals) in the model. A model with less variability in the error terms will produce narrower and therefore more useful confidence and prediction intervals.
3. The number of explanatory variables gives some indication of the data requirements of the model. Adding an explanatory variable may reduce the width of confidence and prediction intervals, if it reduces s_ε. However, adding an explanatory variable will also decrease the degrees of freedom for the *t*-score ($n - (k + 1)$) used in the confidence and prediction intervals (each additional variable increases k by 1), and *t*-scores with smaller degrees of freedom are larger (look at the table of *t*-scores in Appendix 3 on page 581 if you are not sure of this). The larger *t*-score may at least partially offset any reduction in s_ε that results from adding an explanatory variable.

The **Multiple Regression Tools** add-in allows you to easily create all possible regression models from a data set. Exhibit 14.20 opposite shows the **Multiple Regression Tools** dialogue box, with the correct choice highlighted: **All Possible Regressions Calculations**.

EXHIBIT 14.20

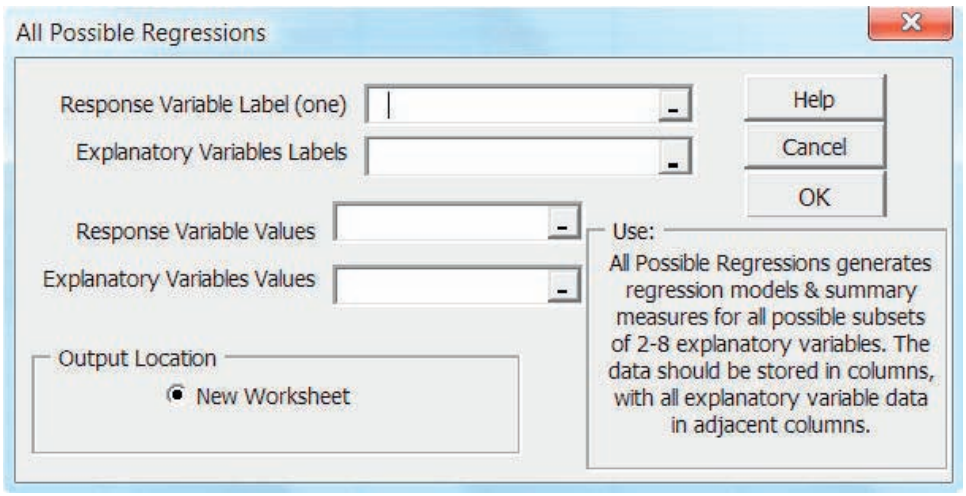
Multiple Regression Tools Add-In, All Possible Regressions Calculations



With this choice selected, click **OK**, and the next dialogue box will be as shown in Exhibit 14.21.

EXHIBIT 14.21

All Possible Regressions Dialogue Box



You are required to:

1. Input the locations of the response (y) and explanatory (x) variable labels (Sales for the response variable for the Woodbon data, and Mortgage Rates, Housing Starts, and Advertising Expenditure for the explanatory variables).
2. Input the locations of the response (y) and explanatory (x) variable values.

The output is illustrated in Example 14.5A on the next page.

EXAMPLE 14.5A

Assess all possible regressions

Use Excel to create all possible regression models for Woodbon sales, using housing starts (which we will refer to as x_1), mortgage rates (x_2), and advertising expenditure (x_3) as possible explanatory variables.

The output from the **All Possible Regressions Calculations** results are shown below in Exhibits 14.22a and b. The output is quite long, and when you are working in Excel, you will have to scroll up and down to see everything.

EXHIBIT 14.22

Results of All Possible Regressions for Woodbon Sales Model, with Housing Starts, Mortgage Rates, and Advertising Expenditure as Explanatory Variables.

a)

Multiple Regression Tools-All Possible Models - Calculations				
Model Number	Adjusted R^2	Standard Error	K	Significance F
1	0.816696238	11651.4898	1	2.73738E-11
Variable Labels	Coefficients	p-value		
Intercept	140617.432	1.92382E-17		
Mortgage Rates	-7201.80514	2.73738E-11		
Model Number	Adjusted R^2	Standard Error	K	Significance F
2	0.038514112	26685.00127	1	0.161030209
Variable Labels	Coefficients	p-value		
Intercept	35847.59546	0.129611207		
Housing Starts	9.686480483	0.161030209		
Model Number	Adjusted R^2	Standard Error	K	Significance F
3	0.881451735	9370.081561	1	9.1661E-14
Variable Labels	Coefficients	p-value		
Intercept	6291.657683	0.191956707		
Advertising Expenditure	35.16671582	9.1661E-14		
Model Number	Adjusted R^2	Standard Error	K	Significance F
4	0.824051049	11415.34684	2	1.41103E-10
Variable Labels	Coefficients	p-value		
Intercept	160079.6411	9.65112E-11		
Mortgage Rates	-7626.20345	6.38131E-11		
Housing Starts	-4.56439049	0.160992622		

b)

Model Number	Adjusted R^2	Standard Error	K	Significance F
5	0.925576239	7424.23274	2	3.01022E-15
Variable Labels	Coefficients	p-value		
Intercept	59670.88304	0.000196457		
Mortgage Rates	-3132.530105	0.000433979		
Advertising Expenditure	22.74342288	1.54935E-06		
Model Number	Adjusted R^2	Standard Error	K	Significance F
6	0.886993562	9148.446728	2	5.57177E-13
Variable Labels	Coefficients	p-value		
Intercept	16136.21269	0.053896821		
Housing Starts	-3.761645082	0.144006697		
Advertising Expenditure	36.68747295	2.43629E-13		
Model Number	Adjusted R^2	Standard Error	K	Significance F
7	0.944400501	6416.987757	3	8.35857E-16
Variable Labels	Coefficients	p-value		
Intercept	80640.56739	4.30411E-06		
Mortgage Rates	-3521.914719	2.64897E-05		
Housing Starts	-5.477255203	0.005171979		
Advertising Expenditure	23.41350711	1.15414E-07		

So which model is best? While the results shown in Exhibit 14.22[6] can help us assess the various regression models, these values cannot tell the whole story. There is a trade-off when variables are added to the model. While the model may have more explanatory power, it will also be more complicated, and more difficult to maintain.

If we look at Exhibit 14.22, we see that the model with all three explanatory variables has the highest adjusted R^2 and the lowest standard error. However, the model has a negative coefficient for housing starts, which does not seem reasonable. As well, adding the housing starts variable only slightly improves it from the model with just advertising expenditure and mortgage rates.

Whichever model we choose, it is important to check that it meets the required conditions described in Section 14.2. An F-test of the significance of the model should also be conducted. While it would be disappointing if the "best" model from all of the possible regressions did not meet the required conditions, it would not be legitimate to use the model for prediction or confidence intervals if it did not.

By now you can see that building a multiple regression model is an iterative process. It can take some time to build, assess, and ultimately decide on the preferred model. Fortunately, it is fairly easy to explore the possibilities with software such as Excel.

There is one more consideration that is important as we build multiple regression models. Whenever we use more than one explanatory variable, we have to consider that there may be interactions among these variables.

[6] If your output for Model 5 does not match what is shown in this exhibit, see the note called "Excel's Floating Point Problem" at the end of the chapter, after the Chapter Review Exercises.

A New Consideration: Multicollinearity

Collinearity occurs when two of the explanatory variables are related to each other. Multicollinearity occurs when more than two of the explanatory variables are related to each other. Generally, this potential problem with multiple regression is referred to as "multicollinearity."

Multicollinearity may cause one or more of the following problems:

1. The adjusted R^2 is large and the F-test shows the overall model is significant, but one or more of the estimated regression coefficients are statistically insignificant.
2. The estimated regression coefficients are not stable. The values change significantly when explanatory variables are added to the regression relationship.
3. The estimated regression coefficients do not make sense. They are larger or smaller than would seem appropriate, or they have an unexpected sign.

Because of these problems, it is important to consider multicollinearity when building a multiple regression model. Some degree of multicollinearity is present in almost every multiple regression model.

The first method of guarding against multicollinearity is to choose the explanatory variables carefully. If mortgage rates, for instance, are being considered as an explanatory variable in the Woodbon model, it would not make sense to include prime rates or another mortgage rate in the model. Because such variables are likely highly correlated with each other, most of the explanatory power is gained when the first variable is introduced. The second variable will not likely tell us more about the response variable.

There are various methods aimed at identifying collinear variables if the relationship between them is not immediately obvious. One of these is to create a scatter diagram of the relationship of every explanatory variable with every other explanatory variable.

In the Woodbon model, if we are thinking of choosing the model with advertising expenditure and mortgage rates, we could create a scatter diagram of advertising expenditure and mortgage rates. Such a scatter diagram is illustrated in Exhibit 14.23 opposite. It appears there is a fairly strong negative correlation between the two variables.

Another method of assessing the correlation between the explanatory variables is to create a correlation matrix for the variables. This is easy to do with Excel, using the **Correlation** tool of **Data Analysis**. If you select the adjacent columns of data, Excel will produce a correlation matrix such as the one shown for the Woodbon problem in Exhibit 14.24. Note that it is helpful to select the labels along with the data when using this Excel tool.

The correlation coefficients tell us something about how the variables are related as pairs. Of course, it is also possible that one explanatory variable could be simultaneously related to two other explanatory variables, and neither the scatter diagrams nor the correlation matrix will reveal this. The scatter diagrams and the correlation matrix will help identify obvious pair-wise correlations between variables, but other harder-to-identify sources of multicollinearity may also be present.

Whenever the correlation coefficients are close to 1 or –1, there are potential problems with multicollinearity. For example, in Exhibit 14.24 opposite, we see a correlation coefficient of 0.946 between advertising expenditure and the leading indicator. If we had not already eliminated the leading indicator as a potential explanatory variable, we would probably have eliminated it because of its high correlation with advertising expenditure.

EXHIBIT 14.23

Scatter Diagram of Mortgage Rates and Advertising Expenditure

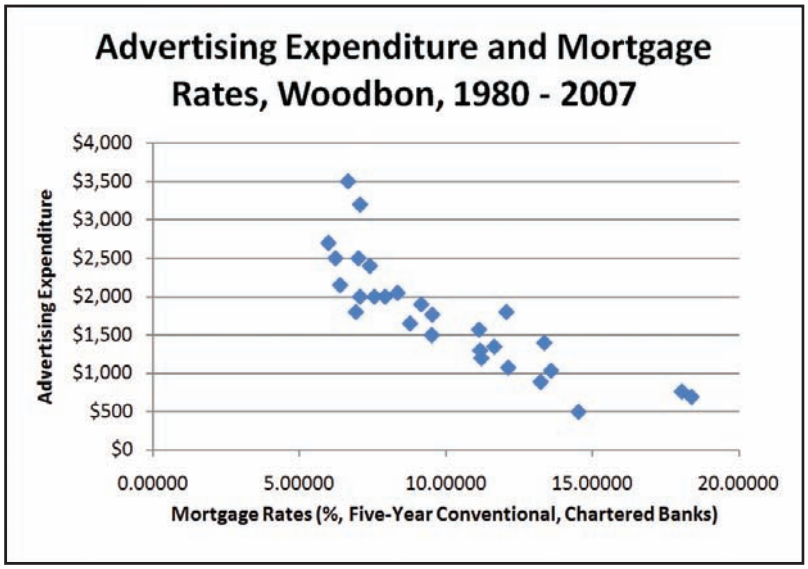


EXHIBIT 14.24

Correlation Matrix for Woodbon Variables

	Mortgage Rates	*Housing Starts*	*Advertising Expenditure*	*Leading Indicator*	*Sales*
Mortgage Rates	1.0000				
Housing Starts	-0.4168	1.0000			
Advertising Expenditure	-0.8424	0.3850	1.0000		
Leading Indicator	-0.7598	0.4473	0.9460	1.0000	
Sales	-0.9075	0.2723	0.9412	0.8811	1.0000

Including the leading indicator variable in the model would have robbed advertising expenditure of its explanatory power.

The collinearity between advertising expenditure and the leading indicator is unexpected. There is no obvious reason for the two variables being connected. However, the mathematical connection is clear, and because of it we should not include both variables in the model.

The correlation coefficient between mortgage rates and advertising expenditure is –0.8424, confirming what we saw in the scatter diagram: there is a fairly strong negative relationship between the two variables. Does this mean we should reject this regression model?

The answer depends on how Kate Cameron intends to use the model. Because of the collinearity between mortgage rates and advertising expenditure, she should be careful interpreting the regression coefficients. In the Woodbon sales model that includes mortgage rates and advertising expenditure, $y = \$59{,}670.88 - \$3{,}132.53x_2 + 22.74x_3$, the x_3 coefficient is 22.74. Normally, we would interpret this coefficient as follows: for a given

mortgage rate, each additional dollar spent on advertising increases Woodbon sales by $22.74. However, because of the collinearity, such an interpretation is not reliable. If Kate wants to know the true nature of the relationship between Woodbon sales and advertising, she should not rely on a model that includes both advertising expenditure and mortgage rates. However, if Kate's only purpose is to predict Woodbon's sales, this model is still probably the best, because it has a high adjusted R^2 and a fairly low standard error.

Example 14.5B below discusses a case in which there is strong collinearity between explanatory variables, and suggests a way to deal with the problem to improve the regression model.

EXAMPLE 14.5B

Multiple regression, dealing with collinearity

EXA14-5b 

A sales manager is trying to build a model to predict sales by region. He has data on population and total income in each region. The manager begins by creating a regression model including both total income and population as explanatory variables. An excerpt from the Excel output is shown below in Exhibit 14.25. Examine the output and explain the results.

EXHIBIT 14.25

Regression Output for Example 14.5B

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.722
R Square	0.521
Adjusted R Square	0.485
Standard Error	686.664
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	13833194.713	6916597.356	14.669	0.000
Residual	27	12730684.665	471506.839		
Total	29	26563879.378			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	26453.443	582.730	45.396	0.000	25257.781
Total Income (Millions)	0.592	0.783	0.757	0.456	-1.014
Population	0.040	0.034	1.174	0.251	-0.030

The regression model has an adjusted R^2 value of 0.485. The model is significant, according to the F-test (the *p*-value is approximately zero). However, the *p*-values for each of the explanatory variables are quite large, indicating that none of them is significant. Since these are the only variables in the model, this does not make sense.

Total income for a region will of course be closely related to the population of the region. Regions with larger populations will have greater total incomes. In this data set, the correlation coefficient between total income and population is 0.936. If the sales manager wants to investigate the effect of income on sales, he should take out the population effect by working with per capita income. Adjusting the income data and exploring the new regression model are left to the reader as an exercise (see Develop Your Skills 14.5, Exercise 22).

DEVELOP YOUR SKILLS 14.5

MyStatLab

The Develop Your Skills exercises in this chapter frequently refer to a data set called "Salaries."

21. Does the Woodbon model that includes mortgage rates and advertising expenditure meet the required conditions for regression? If so, conduct an F-test on the significance of the model.

SEC14-2

22. Adjust the total income data for Example 14.5B, and analyze the new regression model that includes both per capita income and population. Is the model significant? Are both explanatory variables significant? Continue your analysis of the data and decide on the best regression model for this data set.

EXA14-5b

23. Create a correlation matrix for the variables in the Salaries data set. Discuss which explanatory variables should not be used simultaneously, and which look most promising to explain salaries.

SALARIES

24. Create all other possible regression models for the Salaries data. The models will be based on the following explanatory variables:

a. years of postsecondary education alone
b. years of experience alone
c. age alone
d. age and years of experience

25. Compare all possible models for the Salaries data, and select the "best" regression model.

14.6 USING INDICATOR VARIABLES IN MULTIPLE REGRESSION

The regression analysis discussed so far has investigated quantitative explanatory variables. Sometimes we are interested in the effect that a qualitative characteristic might have on a response variable (for example, male/female, urban/rural). It is possible to include such information in regression analysis with the use of indicator variables (sometimes called "dummy" variables). If the qualitative variable we are interested in is binary, that is, it has only two categories, then we can represent it with a single indicator variable (for example, "0" for male, "1" for female).

Indicator Variables for Qualitative Explanatory Variables with Only Two Categories

Once the qualitative variable is coded, it can be treated like any other variable in the regression analysis. For example, suppose we have data on the income and gender of the head of household for a random sample of credit card holders, as well as the monthly credit card bill.

SEC14-6

The Excel **Regression** output for the data set, with both explanatory variables, is shown on the next page in Exhibit 14.26.

EXHIBIT 14.26

Excel Regression Output for Credit Card Data Set

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.6793385				
R Square	0.46150079				
Adjusted R Square	0.42784459				
Standard Error	394.818505				
Observations	35				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	4274963.023	2137482	13.7122	4.9999E-05
Residual	32	4988212.862	155882		
Total	34	9263175.886			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	581.641159	427.1660441	1.36163	0.18283	-288.467596
Income (000)	18.6627831	5.319860768	3.50813	0.00136	7.82658133
Gender of Head of Household (0=Male, 1=Female)	-341.71492	142.7877078	-2.39317	0.02274	-632.563965

From the F statistic and the associated *p*-value, we can see that the model is significant. There appears to be a relationship between the monthly credit card bill and at least one of income and the gender of the head of household. As well, the *p*-value for the *t*-test of the significance of the gender variable is 0.023, indicating that it is significant in the model (for a significance level of 0.05, for instance).

The regression relationship is as follows (with some rounding of the coefficients):

$$\text{Monthly credit card bill} = \$582 + \$19\ (\text{income in } \$000) - \$342\ (\text{gender variable})$$

Notice that this means the following:

- If the credit card holder is male (gender variable = 0), the regression relationship is:

$$\text{Monthly credit card bill} = \$582 + \$19\ (\text{income in } \$000)$$

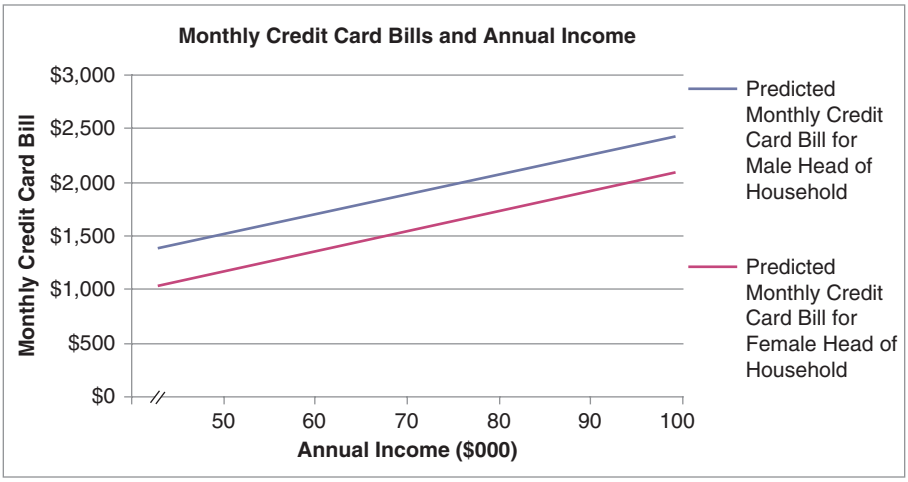
- If the credit card holder is female (gender variable = 1), the regression relationship is:

$$\begin{aligned}\text{Monthly credit card bill} &= \$582 + \$19\ (\text{income in } \$000) - \$342\\ &= (\$582 - \$342) + \$19\ (\text{income in } \$000)\\ &= \$240 + \$19\ (\text{income in } \$000)\end{aligned}$$

A binary indicator variable is sometimes called a "shift" variable, because it shifts the *y*-intercept while leaving the slope unchanged. The two regression relationships are shown in Exhibit 14.27 opposite. The regression relationship predicting monthly credit card bills for

EXHIBIT 14.27

Effect of Gender Variable on Credit Card–Income Relationship



male heads of household (the blue line in the graph) is $342 higher than the regression relationship predicting monthly credit card bills for female heads of household (the red line).

Another way to think of the binary variable is as an on-off switch. When the switch is on (gender variable = 1), we are estimating the monthly credit card bills of female heads of household. When the switch is off (gender variable = 0), we are estimating the monthly credit card bills of male heads of household.

Adding gender to the regression model improves it over the model with income alone (in Develop Your Skills 14.6, Exercise 26, you will get a chance to explore this). However, we should not automatically conclude that gender is the cause of the difference in monthly credit card bills. Other factors such as wealth and the number of people in the household may be the cause of the differences we have observed in the credit card bills of male and female heads of household.

Since the end result of this regression analysis is two regression lines, one for each gender, you might wonder if it would be reasonable to skip the indicator variable, and instead build two regression models, one for male heads of household and one for female heads of household. The advantage of using the indicator variable approach is straightforward. Using the indicator variable in the model allows us to test whether gender has a significant effect on the credit card bills. If we simply build two models, we cannot conduct this statistical test. As well, pooling all of the data together allows us to make better estimates of the coefficients.

Another question you might have is this: is it possible to do regression analysis with only indicator variables, and no quantitative explanatory variables? The answer is yes, and the equivalent procedure is covered in Chapter 11. There, we tested to see if a quantitative response variable (such as battery life) varied according to levels of a qualitative explanatory factor (such as battery brand). The equivalence between the ANOVA procedures in Chapter 11 and regression with indicator variables is illustrated on the next page in Example 14.6. Since our ANOVA analysis considered factors with more than two levels, we will first consider regression using qualitative explanatory variables with more than two categories.

Indicator Variables for Qualitative Explanatory Variables with More Than Two Categories Sometimes the qualitative variable we are interested in has more than two possible (mutually exclusive) results. For example, we might be interested in whether three different brands of battery are associated with different battery life in minutes.

In such a case, we can use a series of indicator variables that tell us about the presence (value = 1) or absence (value = 0) of the brand characteristic. In the case of three battery brands, we will use two indicator variables in combination. For example, we could set the Onever variable = 1 if the brand is Onever, and 0 otherwise. We would set the Durible variable = 1 if the brand is Durible, 0 otherwise. This is sufficient, because if both indicator variables are equal to zero, this necessarily means that the third brand category (PlusEnergy) must apply. Exhibit 14.28 below illustrates. Note that we could have used any of the possible categories as the "missing" one.

EXHIBIT 14.28

Two Indicator Variables for Three Battery Brands

	Two Indicator Variables (in Combination) to Convey Three Battery Brands	
Onever	1	0
Durible	0	1
PlusEnergy	0	0

It is important to use one fewer indicator variables than categories, to avoid problems with the regression analysis. We know that the value of the PlusEnergy indicator variable would be equal to (1 – sum of the values of the indicator variables for the other battery brands). Including a third indicator variable in the model would violate the requirement for independence of the explanatory variables. In addition, it would cause an error in Excel.

Recognize that some common sense is required when introducing qualitative variables into regression analysis. If a qualitative variable has many possible categories, requiring the use of several indicator variables, and if more than one qualitative variable is being considered, the number of explanatory variables can quickly get very high, perhaps too high to be reasonable for small data sets. As always, think carefully before you introduce a qualitative variable (or any variable) into the analysis.

Example 14.6 below illustrates the use of indicator variables for a qualitative explanatory variable with more than two possible categories.

EXAMPLE 14.6

Regression with a qualitative explanatory variable

EXA14-6

The owner of a winery is wondering whether the average purchase of visitors to her winery differs according to age. She asks the cashiers to keep track of a random sample of purchases by customers in three age groups: under 30, 30–50, over 50. Because there is no good reason to ask a customer his or her age, the cashiers guess which age group a customer belongs to (and if they do not guess accurately, the research may not be helpful). Eventually, data from about 50 purchases made by customers in each of the age groups is collected. Does it appear that there is a relationship between customer age and the value of the winery purchase?

First, the winery data must be arranged with all purchases in one column, and the indicator variables set up to indicate age group. Exhibit 14.29 below provides an excerpt from an Excel spreadsheet where the data are set up as required.

EXHIBIT 14.29

Excerpt of Winery Purchase Data Set with Indicator Variables for Age Group

Winery Purchase	1 = Under 30, 0 = Not Under 30	1 = 30-50, 0 = Not 30-50
$ 100.97	1	0
$ 97.55	1	0
$ 134.32	1	0
⋮		
$ 101.96	0	1
$ 75.96	0	1
$ 125.86	0	1
$ 155.44	0	1
$ 125.70	0	1
⋮		
$ 101.79	0	0
$ 168.45	0	0
$ 124.84	0	0

Next, run Excel's **Regression** tool on the data set. The output is as shown in Exhibit 14.30 on the next page.

How do we interpret the regression equation?

- When the "under 30" indicator variable = 1, then:

  Winery purchase for customers under 30 = $132.47 − $54.90 (1) − $12.80 (0) = $77.57

- When the "30–50" indicator variable = 1, then:

  Winery purchase for customers aged 30–50 = $132.47 − $54.90 (0) − $12.80 (1) = $119.67

- When the "under 30" indicator variable = 0, and the "30–50" indicator variable = 0, then the age group is "over 50."

  Winery purchase for customers over 50 = $132.47 – $54.90 (0) – $12.80 (0) = $132.47

Notice that a hypothesis test about the overall regression would report an F statistic of 67.49, with a very low significance level. Thus we can conclude that there is evidence of a relationship between winery purchase and age group. As well, each of the indicator variables included in the regression is significant (at the 5% level of significance).

For comparison, the Excel output for **Anova: Single Factor** is shown in Exhibit 14.31 on the next page.

The F-statistic for the hypothesis test about equality of population means is also 67.49, with the same significance level as in the **Regression** output. Notice that both procedures analyze the data in a similar way. In the regression analysis, the total variation in the *y*-variable (winery purchases) is partitioned into the portion that can be explained by the *x*-variable (the age groups), and the portion that is unexplained (the residuals, or errors). In the

EXHIBIT 14.30

Regression Output for Winery Purchase Data Set

SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.69186325				
R Square	0.47867476				
Adjusted R Square	0.4715819				
Standard Error	24.7237335				
Observations	150				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	82504.42097	41252.2	67.4868	1.6125E-21
Residual	147	89855.6606	611.263		
Total	149	172360.0816			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	132.4674	3.49646392	37.8861	9.7E-78	125.557572
1 = Under 30, 0 = Not Under 30	-54.899	4.944746696	-11.1025	3.5E-21	-64.670973
1 = 30-50, 0 = Not 30-50	-12.7966	4.944746696	-2.58792	0.01062	-22.568573

EXHIBIT 14.31

Anova: Single Factor Excel Output for Winery Purchase Data Set

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Under 30	50	3878.42	77.5684	652.9145		
30-50	50	5983.54	119.6708	555.0899		
Over 50	50	6623.37	132.4674	625.7846		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	82504.42	2	41252.21	67.48684	1.61248E-21	3.057621
Within Groups	89855.66	147	611.263			
Total	172360.1	149				

analysis of variance, the F statistic compares the variation associated with the age groups (the "between groups" variation) with the unexplained variation ("within groups"). Of course, the sums of squares and the mean squares are exactly the same for the two procedures, because they are accomplishing the same tasks. This explains why some of the regression output has the "ANOVA" heading. The Regression sum of squares (SS) in the **Regression** output is the same as the Between Groups SS in the ANOVA output. The Residual SS in the **Regression** output is the same as the Within Groups SS in the ANOVA output.

DEVELOP YOUR SKILLS 14.6

MyStatLab

26. Examine the data set for credit card bills. Create a regression model for credit card bills based on income. Compare this with the regression model for credit card bills based on income and the gender of the head of household. Does adding the gender variable improve the model significantly? SEC14-6

27. Build a regression model, with indicator variables for battery brand, to assess whether there is a significant relationship between battery life in minutes and battery brand. This revisits the battery example in Chapter 11. DYS14-27

28. A sales manager is trying to build a sales forecasting model based on number of sales contacts and region. Is region a significant explanatory variable in this model? DYS14-28

29. A production manager has collected data on the number of units produced and the number of employees at work, for the day shift and the night shift. Is shift a significant explanatory variable for the number of units produced? DYS14-29

30. Statistics Canada collects census data about Canadians every five years. The department provides data files that contain a representative sample of anonymous individual responses to census surveys. A subset of these data is provided in the file DYS14-30. There is information on Canadians from Alberta and Ontario. Age and wages and salaries are shown for individuals who had non-zero wages and salaries in the data set.[7] Use an indicator variable for province, and build a regression model for wages and salaries, based on age and province. Is province a significant explanatory variable in this model? DYS14-30

14.7 MORE ADVANCED MODELLING

This chapter has been an introduction to building mathematical models of linear relationships between quantitative response variables and two or more explanatory variables. Within the chapter we have seen many modelling possibilities.

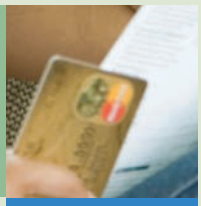
You should be aware that more advanced mathematical modelling techniques exist, which are beyond the scope of this text. It is possible to build models that are polynomial, to account for curvature in the relationships. There are special techniques for time-series trend analysis. With the appropriate training and good computer software, it is possible to build complex and sophisticated models of relationships. However, complex models are not necessarily the "best" models. The simplest model that provides useful predictions is preferred.

Finally, always remember that mathematical models generally cannot prove cause and effect, and we should always be careful in interpreting the results of model building. Even if a model appears to work very well, the true cause-and-effect relationship may not have been revealed.

[7] Data for this exercise are a subset of the data available in the StatsCan microfile. Only age, wages, and salaries for those with non-zero wages and salaries data are used, for only Alberta and Ontario.

Chapter Summary

14



Determining the Relationship

Begin by thinking carefully about the explanatory variables that might reasonably be expected to affect the response variable. Create scatter plots to examine the relationship between the response variable and each explanatory variable. Use Excel's **Regression** tool to estimate the coefficients of the multiple regression relationship.

Checking the Required Conditions

Theoretically, there is a normal distribution of possible y-values for every combination of x-values. The population relationship we are trying to model is as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

We cannot reliably make predictions with the regression equation, or conduct a hypothesis test about the significance of the regression relationship, unless certain conditions are met (these are summarized in the box on page 532). The Guide to Technique: Checking Requirements for the Linear Multiple Regression Model on page 540 outlines a process for checking the required conditions for the regression model.

How Good Is the Regression?

If the required conditions are met, conduct an F-test of the significance of the relationship. This will be of the form:

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1: At least one of the β_i's is not zero.

The F statistic is

$$F = \frac{\dfrac{SSR}{k}}{\dfrac{SSE}{n - (k + 1)}} = \frac{MSR}{MSE}$$

with $(k, n-(k+1))$ degrees of freedom, where n is the number of observed data points and k is the number of explanatory variables in the model. When the response variable is related to at least one of the explanatory variables, MSR will be significantly larger than MSE.

The output of Excel's **Regression** tool provides the F statistic and the p-value for this test. See page 544 for instructions on how to read the output.

If the results of the F-test show that the model is significant, t-tests of the significance of the individual explanatory variables can be conducted. The test of the coefficient of explanatory variable i is conducted as follows.

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$

The test statistic is $t = \dfrac{b_i}{s_{b_i}}$, with $(n - (k + 1))$ degrees of freedom.

The output of Excel's **Regression** tool provides the p-values for the two-tailed tests of significance for the individual coefficients. See page 545 for instructions on how to read the output.

The adjusted R^2 value is a measure of the strength of the relationship between the explanatory variables and the response variable.

$$\text{Adjusted } R^2 = 1 - \frac{\dfrac{SSE}{n - (k + 1)}}{\dfrac{SST}{n - 1}}$$

The adjusted R^2 is calculated by Excel as part of the **Regression** output.

Making Predictions

Two types of estimation intervals can be created if the requirements are met. A regression prediction interval predicts a particular value of y, given a specific set of x-values. A regression confidence interval predicts the average y, given a specific set of x-values. The **Multiple Regression Tools** Excel add-in (**Prediction and Confidence Intervals Calculations**) calculates these intervals (see page 548). Always remember that it is not legitimate to make predictions outside the range of the sample data.

Selecting the Appropriate Explanatory Variables

The goals of a good regression model include the following:

1. The model should be easy to use. It should be reasonably easy to acquire data for the model's explanatory variables.
2. The model should be reasonable. The coefficients should represent a reasonable cause-and-effect relationship between the response variable and the explanatory variables.
3. The model should make useful and reliable predictions. Prediction and confidence intervals should be reasonably narrow.
4. The model should be stable. It should not be significantly affected by small changes in explanatory variable data.

Use Excel to create all possible regression models for all combinations of possible explanatory variables. The **Multiple Regression Tools** Excel add-in (**All Possible Regressions Calculations**) makes this easy to do. The add-in produces a summary report which shows each regression model, the adjusted R^2 value, the standard error (s_ε), and the number of variables (k) for each model. Use these measures to select a "best" model. Be sure to check the model chosen to see that it meets the required conditions. Review the appropriate p-values to ensure that the overall model is significant, and that the individual explanatory variables are significant.

Multicollinearity occurs when one of the explanatory variables is highly correlated with one or more of the other explanatory variables. This can result in unstable or inaccurate regression coefficients. To guard against this problem, choose explanatory variables carefully. Create scatter diagrams of explanatory variable pairs, and create a correlation matrix for all the variables in the model. If there is a pronounced pattern visible in the scatter diagram or a high correlation for a pair of variables, consider including only one of them in the final model.

Using Indicator Variables in Multiple Regression

The effect of a qualitative characteristic on a response variable (for example, male/female, urban/rural) can be modelled with indicator variables (sometimes called "dummy" variables). If the qualitative variable we are interested in is binary, we can represent it with a single indicator variable (for example, "0" for male, "1" for female). If the qualitative variable has more than two possible (mutually exclusive) results, we can use a series of indicator variables that tell us about the presence (value = 1) or absence (value = 0) of the qualitative characteristic. It is important to use one fewer indicator variables than categories, to avoid problems with the regression analysis.

CHAPTER REVIEW EXERCISES

CREDIT CARD

MARKS

The Chapter Review Exercises allow you to explore two major data sets. One, called "Credit Card," contains data about credit card balances and possible explanatory variables such as income and number of people in the household. Another data set, called "Marks," contains data about the final exam mark and marks on evaluations done during the semester (assignments, tests, and quizzes).

WARM-UP EXERCISES

1. The multiple regression model for monthly credit card balances and the age of the head of household, income (in thousands of dollars), and the value of the home (in thousands of dollars) is described in the Excel output shown below in Exhibit 14.32. Interpret the model.

EXHIBIT 14.32

Excel Output for Monthly Credit Card Balances, Income in Thousands, and the Number of People in the Household

SUMMARY OUTPUT				
Regression Statistics				
Multiple R	0.604771472			
R Square	0.365748533			
Adjusted R Square	0.304369359			
Standard Error	435.3412786			
Observations	35			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	3	3387992.99	1129330.997	5.958837626
Residual	31	5875182.895	189522.0289	
Total	34	9263175.886		
	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	38.35751156	539.3083936	0.07112352	0.94375632
Age of Head of Household	0.992027023	11.58540256	0.085627324	0.932313329
Income (000)	22.03697173	8.899239591	2.47627581	0.018939953
Value of Home (000)	0.375387927	3.933953149	0.095422572	0.924593365

2. Refer to the Excel output shown in Exhibit 14.32 above. Is the overall model significant? You will have to estimate the *p*-value from the tables at the back of the text. Use a 5% level of significance.

3. Refer to the Excel output shown in Exhibit 14.32 above. Test the individual coefficients for significance. Use a 5% level of significance.

4. Given your answers to Exercises 2 and 3, what concerns do you have about the model? A correlation matrix for the values in the model is shown below in Exhibit 14.33 below.

EXHIBIT 14.33

Correlation Matrix for Credit Card Data Set

	Age of Head of Household	*Income (000)*	*Number of People in Household*	*Value of Home (000)*	*Monthly Credit Card Bill (Annual Average)*
Age of Head of Household	1				
Income (000)	0.781920213	1			
Number of People in Household	0.044761895	0.144977942	1		
Value of Home (000)	0.713556435	0.623129703	0.022990811	1	
Monthly Credit Card Bill (Annual Average)	0.48555699	0.604253286	0.618354542	0.393600727	1

THINK AND DECIDE

5. Consider the Marks data set, where the goal is to predict the final exam mark. Possible explanatory variables are the marks on Assignments 1 and 2, Tests 1 and 2, and Quiz marks. The tests and the final exam are written in a classroom, with all computations done manually with a calculator. The quizzes are done with online testing software, and the calculations can be done manually with a calculator or with Excel. Students can attempt the quizzes as many times as they wish before the due date (the quizzes are similar but not the same). The assignments are Excel-based and include a written report on the Excel analysis. Based on this information, which of the evaluations do you think would be the best predictor of the final exam mark, and why?

6. Exhibit 14.34 below shows the correlation matrix for the Marks data set. Is there any concern about multicollinearity? Which explanatory variables seem the most promising?

EXHIBIT 14.34

Correlation Matrix for Marks Data Set

	Assignment #1	*Test #1*	*Assignment #2*	*Test#2*	*Quizzes*	*Final Exam Mark*
Assignment #1	1.0000					
Test #1	0.2453	1.0000				
Assignment #2	0.4619	0.3719	1.0000			
Test#2	0.2970	0.5147	0.3514	1.0000		
Quizzes	0.3254	0.3247	0.4117	0.5865	1.0000	
Final Exam Mark	0.3898	0.5195	0.4914	0.7167	0.4992	1.000

7. Exhibit 14.35 on the next page shows the output of the **All Possible Models Calculations** tool in **Multiple Regression Tools** for all of the Marks models with one explanatory variable. Given these results, is there one model that you would choose as better than the rest? If so, explain why.

EXHIBIT 14.35

All Possible Models Calculations Output for Marks Models with One Explanatory Variable

Multiple Regression Tools-All Possible Models - Calculations				
Model Number	Adjusted R^2	Standard Error	K	Significance F
1	0.14285023	18.87369279	1	9.42336E-05
Variable Labels	Coefficients	p-value		
Intercept	41.9078595	9.25152E-11		
Assignment #1	0.39827912	9.42336E-05		
Model Number	Adjusted R^2	Standard Error	K	Significance F
2	0.26205943	17.51213989	1	6.85999E-08
Variable Labels	Coefficients	p-value		
Intercept	30.1007124	2.93098E-06		
Test #1	0.55058773	6.85999E-08		
Model Number	Adjusted R^2	Standard Error	K	Significance F
3	0.23330345	17.85008461	1	4.27328E-07
Variable Labels	Coefficients	p-value		
Intercept	47.5765866	1.19338E-23		
Assignment #2	0.33971697	4.27328E-07		
Model Number	Adjusted R^2	Standard Error	K	Significance F
4	0.50836095	14.29393191	1	3.18098E-16
Variable Labels	Coefficients	p-value		
Intercept	31.7519266	4.18575E-14		
Test#2	0.60113345	3.18098E-16		
Model Number	Adjusted R^2	Standard Error	K	Significance F
5	0.2411435	17.75858482	1	2.6112E-07
Variable Labels	Coefficients	p-value		
Intercept	49.3510102	2.14567E-27		
Quizzes	0.37275646	2.6112E-07		

THINK AND DECIDE USING EXCEL

CREDIT CARD 

8. Use the model for credit card balances illustrated in Exhibit 14.32 to create a 95% prediction interval for the monthly credit card balance of a credit card holder where the age of the head of household is 45, income is $65,000, and the value of the home is $175,000. Do you think this regression model is useful?

MARKS 

9. Use Excel to create all possible Marks models, and then consider those that have two explanatory variables. Note that there are 10 such models. Which of these models is best, and why? Is this model a real improvement on the best single-variable model? Explain.

10. Check that the best model you selected in Exercise 9 meets the required conditions.

11. For the Marks data set, create and examine all models with three explanatory variables that include the mark on Test 2. Note that there will be six of these models. Does any of these represent a real improvement on the best two-variable model you selected in Exercise 9? Explain.

12. Create a regression model for the Marks data using all of the explanatory variables. In light of the work you did in Exercises 9, 10, and 11, is this the best model? Explain.

13. Use the model you decided was best for the Marks data to predict the final exam mark of a student who received a mark of 55 on Assignment 1, 60 on Test 1, 65 on Assignment 2, 70 on Test 2, and 95 on the quizzes.

14. A researcher has collected a random sample of data about Honda Accords for sale in Ontario. The data indicate year of the car, number of kilometres, and list price. Create and analyze all possible regression models for these data. Be sure to check that the required conditions are met. Is your model useful in terms of predicting the list price of used Honda Accords?

 CRE14-14

15. Think about your analysis in Exercise 14. Is the year of the car a quantitative variable? Create an indicator variable for the year of the car, and rebuild the model. Describe the models, and choose the best one.

16. A chain of retail outlets famous for their delicious (if unhealthy) doughnuts is looking for a new location. The company is trying to use data on local median income, population in the local area, and traffic flows by a proposed location to decide where to open a new store. The company has collected data for a number of existing stores. Investigate these data, and make a recommendation to the company about how to proceed.

 CRE14-16

17. A researcher has discovered some extra data for the doughnut store location decision described in Exercise 16 above. Information was collected about whether each location was within a five-minute drive of a major highway (1 = within a five-minute drive of a major highway, 0 = otherwise). Re-analyze the data, including this extra information.

 CRE14-17

18. An MBA (Master of Business Administration) student decides to see if he can predict the Standard and Poor's Toronto Stock Exchange Composite Index from the price of one or more share prices of Canadian companies.

a. The student collects monthly historical price data (November 2002 to November 2008) for stocks from some important Canadian sectors[8]:
   - Rona Incorporated, the largest Canadian distributor and retailer of hardware, home renovation, and gardening products
   - Royal Bank of Canada, a major Canadian bank
   - Petro Canada, a Canadian oil and gas company with international interests
   - Potash Corporation of Saskatchewan, an integrated producer of fertilizer, industrial, and animal feed products

   Do any of these stocks (or a combination of these stocks) provide a good predictor of the TXS Composite Index?

b. During the fall of 2008, the world economy experienced an unprecedented crisis and stock markets around the world gyrated wildly. Is there evidence of this in the data you examined in part a of this question? Would it be wise to try to build a model to predict the TSX Composite Index using data from this period? Explain.

 CRE14-18

19. Statistics is a course with a bad reputation. Students tend to expect that they will have difficulty with the course, even when they do not know exactly what the course is about. A student decides that he wants to place Statistics in a proper context, and he collects data on a random sample of students studying in their third semester (the beginning of second year).

 CRE14-19

[8] Yahoo! Finance Canada, "Potash Corp Sask Com NPV (POT.TO), Royal Bk of Canada Com NPV (RY.TO), S&P/TSX Composite index (Interi ^GSPTSE), Rona Inc Com NPV (RON.TO), Petro Canada Com NPV (PCA.TO), Historical Prices, November 20, 2008," **http://ca.finance.yahoo.com**, accessed November 20, 2008.

The student attempts to predict the Statistics mark from the marks in other courses. Help him by deciding which of the possible models is best.

20. Check the required conditions of the model you chose in Exercise 19. If the required conditions are not met, do some further analysis and develop a model that will predict the Statistics mark and meets the required conditions.

21. Use the best model you created in Exercise 20 to predict the Statistics mark of a student who received 65 in all of the other courses.

TEST YOUR KNOWLEDGE

CRE14-22 

22. Marcharpex is a company selling specialized software products to a select number of manufacturing companies. The company relies on senior salespeople to make contacts, sell the product, and provide a company contact for after-sales support. The company is wondering if its sales model is effective, and it has collected some data on

- the years of experience of the salesperson
- the monthly travel and entertainment budget of the salesperson
- the local advertising budget (monthly) for the salesperson's area
- the sales in the area

Analyze the data, and select the best model to predict sales. Be sure to check the required conditions. Once you select the best model, create a 95% prediction interval (approximate) of the sales for a salesperson who has 15 years of experience, a monthly travel and entertainment budget of $2,000, and a local advertising budget of $4,000.

A NOTE ABOUT EXCEL'S FLOATING POINT PROBLEM

Depending on your computer, you may have seen a different result in your output for the Woodbon model with mortgage rates and advertising expenditure. The model might have been $y = \$10.31 - \$0.87x_2 + 0.01x_3$, with an adjusted R^2 of 0.948, and a standard error of only $2! The first time the author ran the regression in Excel (on an older computer) this was the result. However, if we examine this model, we see that it does not make any sense. If advertising expenditure were $3,000 and mortgage rates were 7%, this model predicts that Woodbon's sales would be $\$10.31 - \$0.87(7) + 0.01(\$3{,}000) = \$34.22$. This prediction is clearly unreasonable. This is a good lesson in using some common sense, and not relying too much on measures such as R^2 to choose a regression model. But what went wrong?

Excel has a "floating point" problem that sometimes produces inaccurate results when the data used to calculate the model are on significantly different scales. Advertising expenditures range from $500 to $3,500 and sales range from $21,334 to $115,320. In contrast, mortgage rates range from 5.99 to 18.38. Because mortgage rates vary only by units (as these rates are expressed), and the other variables vary by hundreds or thousands, the scales are not the same general order of magnitude. The alternate model shown above does not make sense, because Excel and an older computer did not successfully handle the situation. While such a problem does not arise often, it can be a good idea to scale your input data to the same order of magnitude. As well, if Excel produces nonsensical results, you should check the scale of your input data and re-scale if necessary.

Fortunately, the fix is easy. If this happened to you, simply change the scale of the mortgage rates by multiplying by 100, for example. A mortgage rate data point of 14.52083 becomes 1452.083, which is effectively $100x_2$. Then the mortgage rate data points will vary by hundreds, and will be on a similar scale with the other explanatory variable in the model. The multiple regression model that includes these adjusted mortgage rate data points and advertising expenditure is:

$$y = \$59{,}670.88 - \$31.3253\,(100x_2) + 22.74x_3.$$

This can of course be rewritten as

$$y = \$59{,}670.88 - \$3{,}132.53x_2 + 22.74x_3.$$

This model matches the output in Exhibit 14.22.