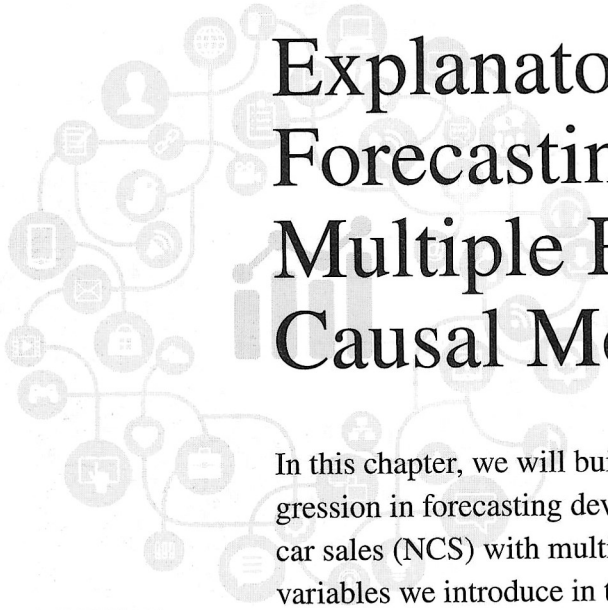
Chapter Five

Explanatory Models 1. Forecasting with Multiple Regression Causal Models



In this chapter, we will build on the introduction to the use of regression in forecasting developed in Chapter 4. We will model new car sales (NCS) with multiple independent variables. One of the variables we introduce in this chapter provides a way in which we can take into account consumer attitudes. To do so, we will use the University of Michigan Index of Consumer Sentiment. We also introduce a new type of independent variable called a "dummy variable." These variables will be used to help account for seasonality in data as well as other events that can influence sales. We will continue with our ongoing example of forecasting The Gap sales at the end of this chapter. These extensions of the bivariate regression model take us into the realm of multiple regression. We will begin by looking at the general multiple-regression model.

LEARNING OBJECTIVES

After studying this chapter, you should be able to:

1. Explain the difference between bivariate (simple) regression and multiple regression.
2. Explain the new (fifth) step for evaluating a multiple regression model.
3. Describe how a regression plane differs from a regression line.
4. Explain what is meant by a "dummy variable."

5. Describe some ways "dummy variables" can be useful in regression models.
6. Explain things that should be considered when selecting independent variables for a multiple regression model that will be used to make a forecast.

THE MULTIPLE-REGRESSION MODEL

Multiple regression is a statistical procedure in which a dependent variable (Y) is modeled as a function of more than one independent variable ($X_1, X_2, X_3, \ldots, X_n$).[1] The population multiple-regression model may be written as:

$$\begin{aligned} Y &= f(X_1, X_2, X_3, \ldots, X_n) \\ &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + \varepsilon \end{aligned}$$

where β_0 is the intercept and the other β_i's are the slope terms associated with the respective independent variables (i.e., the X_i's). In this model, ε represents the population error term, which is the difference between the actual Y and that predicted by the regression model ($\hat{Y}$).

The ordinary least-squares (OLS) criterion for the best multiple-regression model is that the sum of the squares of all the error terms is minimized. That is, we want to minimize $\Sigma\varepsilon^2$, where

$$\Sigma\varepsilon^2 = \Sigma(Y - \hat{Y})^2$$

Thus, the ordinary least-squares criterion for multiple regression is to minimize:

$$\Sigma(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - \cdots - \beta_k X_k)^2$$

The process of achieving this is more complicated than in the bivariate regression case and involves the use of matrix algebra.

Values of the true regression parameters (β_i) are typically estimated from sample data. The resulting sample regression model is:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k$$

where b_0, b_1, b_2, b_3, and so on, are sample statistics that are estimates of the corresponding population parameters β_0, β_1, β_2, β_3, and so on. Deviations between the predicted values based on the sample regression ($\hat{Y}$) and the actual values (Y) of the dependent variable for each observation are called *residuals* (or they are called *errors*) and are equal to ($Y - \hat{Y}$). The values of the sample statistics b_0, b_1, b_2, b_3, and so on, are almost always determined for us by a computer software package. Standard errors, t-ratios, the multiple coefficient of determination, the Durbin-Watson statistic, and other evaluative statistics, as well as a table of residuals, are also found in most regression output.

[1] For more detailed discussions of the multiple-regression model, see the following: John Neter, William Wasserman, and Michael H. Kutner, *Applied Linear Regression Models* (New York: McGraw-Hill, 1996), and Damodar N. Gujarati, *Basic Econometrics* (New York: McGraw-Hill, 2003). The latter is particularly recommended.

INITIAL CONSIDERATIONS WHEN SELECTING INDEPENDENT VARIABLES

As with bivariate regression, the process of building a multiple-regression model begins by identifying the dependent variable.

As with bivariate regression, the process of building a multiple-regression model begins by identifying the dependent variable. In our context, that is the variable that we are most interested in forecasting. It may be some "prime mover" such as disposable personal income or another macroeconomic variable, or it may be total company sales, or sales of a particular product line, or the number of patient-days for a hospital, or state tax revenues.

Once the dependent variable is determined, we begin to think about what factors contribute to its changes. In this chapter, as we will use new car sales (NCS) as the dependent variable we wish to forecast. We will use a measure of income as well as other independent (causal) variables that might improve the model. We want to think of other things that influence NCS but that do not measure the same basic relationship that is being measured by disposable personal income per capita (DPIPC). Think, for example, of the possibility of adding the gross domestic product (GDP) to the model. Both GDP and DPIPC are measures of income in the economy, so there would be a lot of overlap in the part of the variation in NCS they explain. In fact, the correlation between GDP and DPIPC is +0.99. A similar overlap would result if population and DPIPC were used in the same model. There is a high correlation between population size and real disposable personal income per capita (approximately + 0.95), and so they would have a lot of overlap in their ability to explain variations in NCS. Such overlaps can cause a problem known as *multicollinearity,* which we will discuss later in this chapter.[2]

In considering the set of independent variables to use, we should find ones that are not highly correlated with one another.

Thus, in considering the set of independent variables to use, we should find ones that are not highly correlated with one another. For example, suppose that we hypothesize that at least some portion of NCS may be influenced by the unemployment rate. It seems less likely that there would be a stronger correlation between personal income and the unemployment rate than between personal income and either GDP or population size. The correlation between the unemployment rate and disposable personal income turns out to be just 0.15, so there is less overlap between those two variables.

Sometimes it is difficult or even impossible to find a variable that measures exactly what we want to have in our model. For example, in the NCS model, we might like to have as a measure of the interest rate a national average of the rate charged on installment loans. However, a more readily available series, the prime interest rate (PR), may be a reasonable proxy for what we want to measure since all interest rates tend to be closely related.

As we build a model of new car sales (NCS), we will also begin looking at the relationship between NCS and other possible causal variables. In Figure 5.1, you see a plot of NCS. You see a seasonal pattern in the data. Thus, we will want

[2] Note that multicollinearity in regression analysis is really just strong correlation between two or more independent variables. Correlation here is measured, just as we did in Chapter 2 with the Pearson product-moment correlation coefficient.

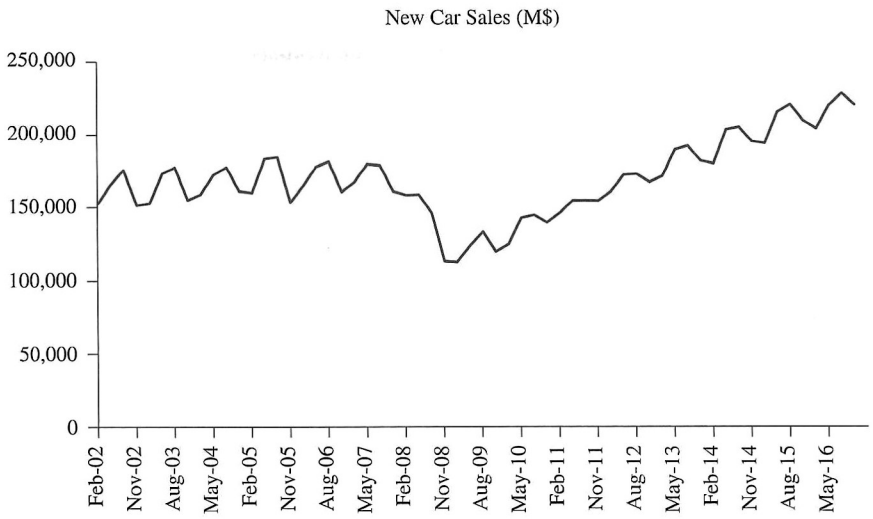


FIGURE 5.1
New Car Sales (NCS) in Millions of Dollars (c5f1)

Source: economagic.com.

to consider adding another variable (or set of variables) to account for the seasonality in NCS. But how do we measure spring, or fall, or summer, or winter? The seasons are qualitative attributes that have no direct quantitative counterpart. We will see (in the section "Accounting for Seasonality in a Multiple-Regression Model") that a special kind of variable, known as a *dummy variable,* can be used to measure such a qualitative attribute as spring.

Something to consider when developing a multiple regression model to use in forecasting is that you must be able to forecast all the independent variables. This suggests that one may want to follow the KIS principle (**K**eep **I**t **S**imple). There may be a trade-off between explanatory power and the number of independent variables used.

DEVELOPING MULTIPLE-REGRESSION MODELS

We will develop several models for NCS building in complexity as we go. When we think about car sales, we are likely to first think about the available purchasing power consumers have at their disposal. For this, we use disposable personal income per capita (DPIPC). Before we consider this and other models for NCS, look carefully at the data in the time series graph in Figure 5.1. The data are quarterly for NCS from 2002Q1 through 2016Q4. You see some seasonality and a big drop during the recession that started in 2008.

Our beginning bivariate regression model is:

$$\text{NCS} = b_0 + b_1(DPIPC)$$
$$\text{NCS} = 66,396.257 + 2.866(DPIPC)$$

where NCS stands for new car sales and DPIPC is disposable personal income per capita (per person). The coefficient for DPIPC is logical, significantly positive at

a 95 percent confidence level (t = 4.144), but the coefficient of determination (R^2) in this case is low (22.8 percent). Also, the DW is 0.283, showing positive serial correlation. So, we must ask what else might affect car sales. We will consider a measure of how people feel about the economy, the unemployment rate, and the prime interest rate.

We will expand the model to include a measure of consumer attitudes, the unemployment rate (UR), and the bank prime rate (PR) as a proxy for all the types of car financing. To capture consumer attitudes about the economy and their place in the economy, we will use the University of Michigan Index of Consumer Sentiment (UMICS). The multiple regression model is:

$$\text{NCS} = b_0 + b_1(\text{DPIPC}) + b_2(UMICS) + b_3(UR) + b_4(PR)$$

Before running the regression, think about what signs should be expected for b_1, b_2, b_3, and b_4. Business and economic logic would suggest that b_1 should be positive ($b_1 > 0$) because the more income people have, the more likely they are to purchase a car. For b_2, we also expect a positive sign since the UMICS increases when people feel better about the state of the economy and are, therefore, more likely to make a major purchase ($b_2 > 0$). When there is high unemployment, we would expect fewer people to be in the market to buy a car, so b_3 should be negative ($b_3 < 0$). If the cost of borrowing decreases, we expect more car sales, and thus, we expect b_4 to be negative ($b_4 < 0$). As shown in Table 5.1, the regression results support this notion. The model is:

$$\text{NCS} = 25{,}304.35 + 3.227(\text{DPIPC}) + 1{,}123.363(UMICS) - 7{,}659.605(UR) - 3{,}216.777(PR)$$

Statistical evaluation of this model, based on the results in Table 5.1, will be considered in the next section. The data are in the c5t1&f2 Excel file. For now, we can see that at least the signs for the coefficients are consistent with our expectations.

The predicted values from this model are plotted in Figure 5.2.

TABLE 5.1
Regression Results for New Car Sales (NCS) as a Function of Disposable Income per Capita (DPIPC), the University of Michigan Index of Consumer Sentiment (UMICS), the Unemployment Rate (UR), and the Prime Interest Rate (PR) (c5t1&f2)

Regression Statistics					
Adjusted *R* Square	0.816				
Standard Error	11,690.644				
Observations	60			DW =	1.376
	Coefficients	Std Error	t Stat	P-value	P/2
Intercept	25,304.350	37,454.530	0.676	0.502	0.251
DPIPC	3.227	0.392	8.240	0.000	0.000
UMICS	1,123.363	210.582	5.335	0.000	0.000
UR	−7,659.605	1,661.586	−4.610	0.000	0.000
PR	−3,216.777	1,412.449	−2.277	0.027	0.013

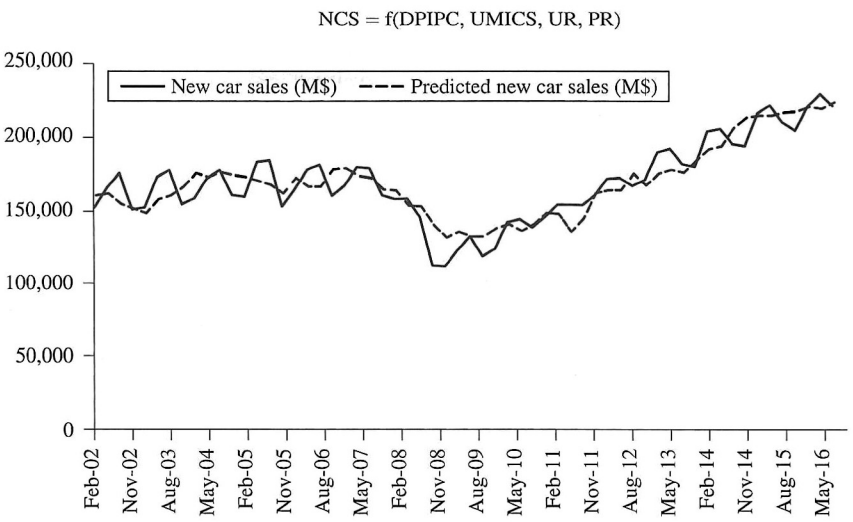


FIGURE 5.2
New Car Sales and Predicted Values (M$) Based on The Model in Table 5.1 (c5t1&f2)

In Figure 5.2, the solid line shows actual values of new car sales (NCS) for 20022Q1 through 2016Q4. The dotted line shows the values predicted by this model for 2002Q1 through 2016Q4. Based on the results in Table 5.1, we will do a statistical analysis of the model. This model is in a five-dimensional space that we cannot visualize. A bivariate regression model is two dimensional, so it is easy to visualize as a line in a graph. If we used only two independent variables, we could construct a three-dimensional graph. Let us briefly look at such a graph for NCS.

A Three-Dimensional Scattergram

Suppose we model NCS as a function of only DPIPC and the UR. The model would be:

$$NCS = 113,524.568 + 3.469(DPIPC) - 10,636.670(UR)$$

We see that the signs for the two independent variables are consistent with business/economic logic.

In our three-variable case (with NCS as the dependent variable and with DPIPC and UR as independent variables), three measured values are made for each sample point (i.e., for each quarter). These observations can be depicted in a scatter diagram like those in Chapter 2, but the scatter diagram must be three -dimensional. Figure 5.3 shows the new car sales (NCS) of any observation as measured vertically from the DPIPC/UR plane. The value of UR is measured along the "UR" axis, and the value of DPIPC is measured along the "DPIPC" axis. All 60 observations are represented as points in the upper diagram. The regression plane is added in the lower panel of Figure 5.3.

In a multiple-regression analysis, our task is to suspend a linear three-dimensional plane (called the *regression plane*) among the observations in such a way that the plane best represents the observations. The multiple-regression analysis

estimates an equation ($Y = a + b_1X + b_2Z$) in such a manner that all the estimates of Y made with the equation fall on or close to the surface of the linear plane.

If all the actual data points were to lie very close to the regression plane, the adjusted R-squared of the equation would be very high.

If all the actual data points were to lie very close to the regression plane, the adjusted R-squared of the equation would be very high. If on the other hand, most of the actual points were far above or below the regression plane, the adjusted R-squared would be lower than it otherwise would be. Normally, regression packages do not have a provision for the graphing of output in three-dimensional form. This is because relatively few of the problems faced in the real world involve exactly three variables. Sometimes you are working with only two variables, while at other times you will be working with more than three. Thus, a three-dimensional diagram will

FIGURE 5.3 New Car Sales (NCS) in Millions of Dollars with DPIPC and UR Viewed in Three Dimensions The Three Dimensional Scattergram Below Shows how Car sales Vary as Both the UR and DPIPC Vary. In Figure 5.3, continued on the next page, you will see the Regression Plane in this Three Dimensional Space. (c5t1&f2)

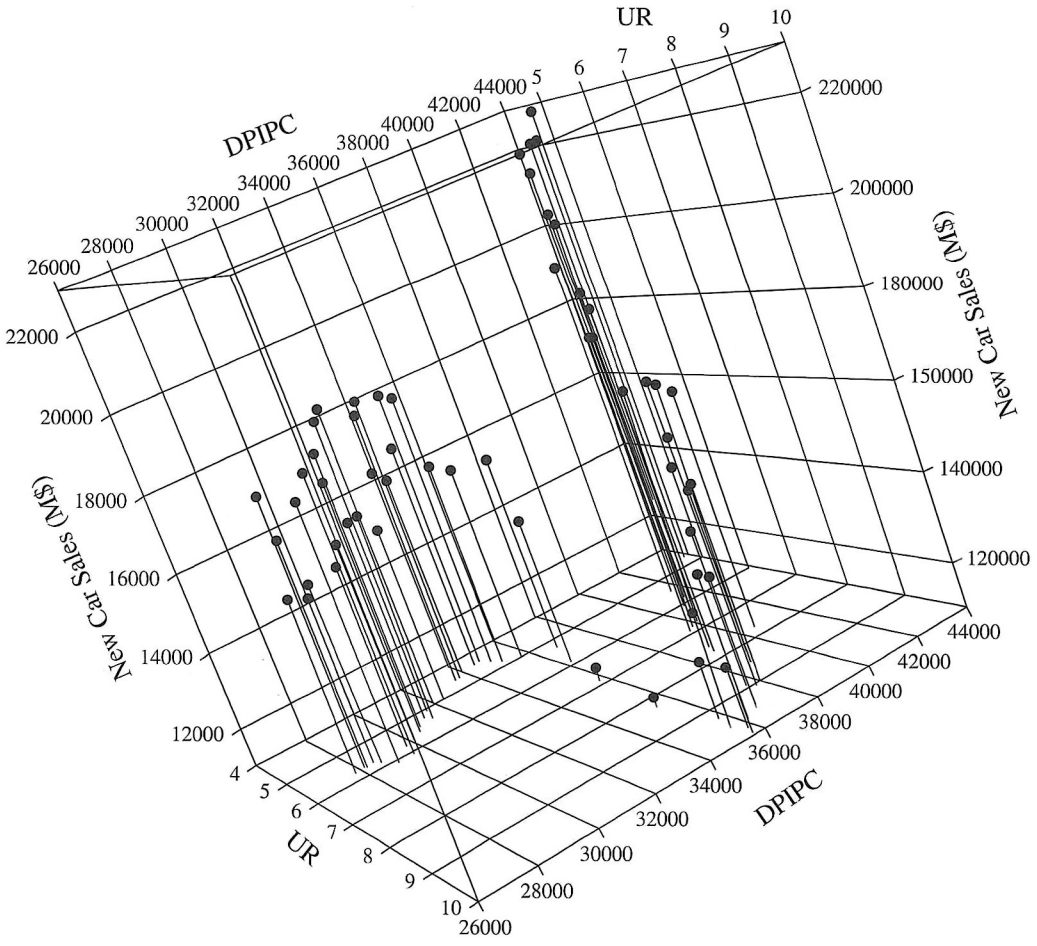
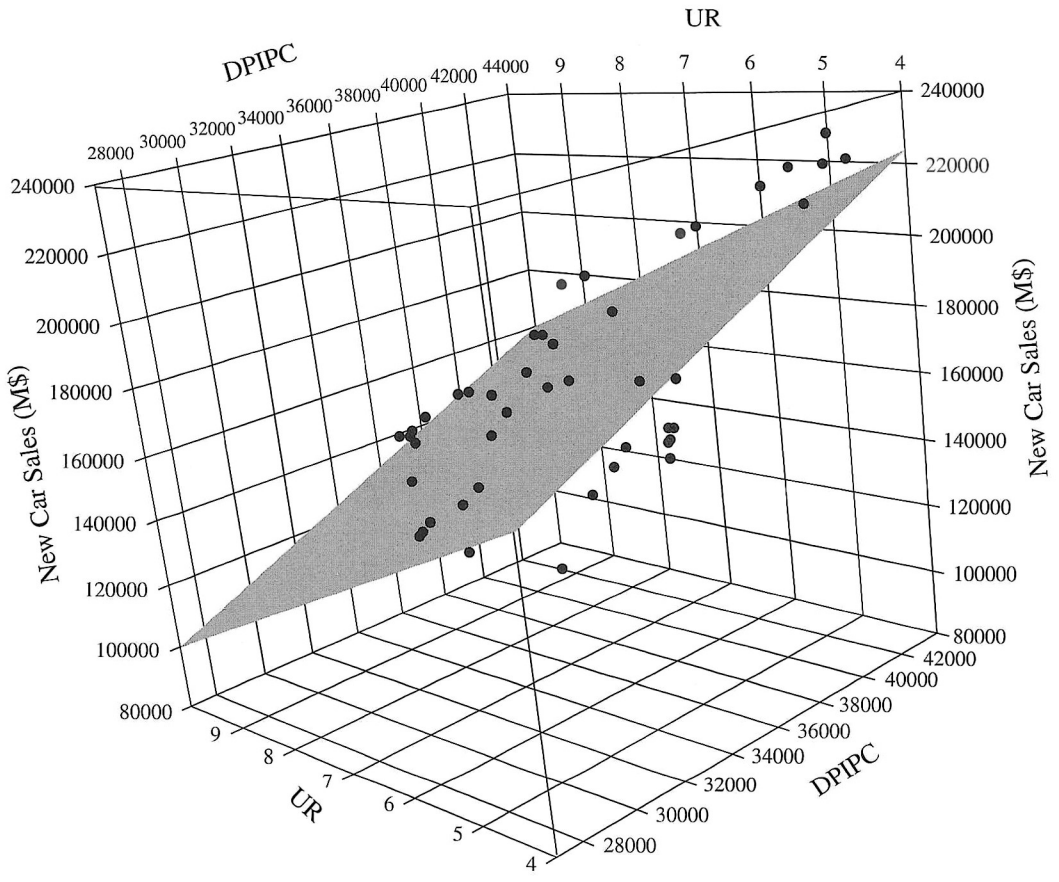


FIGURE 5.3 (continued)



be useful only in a few cases. The plane is a normal plane when there are two independent variables, and it is called a *hyperplane* (more than three-dimensional) when there are more than two independent variables, as in the model shown in Table 5.1.

STATISTICAL EVALUATION OF MULTIPLE-REGRESSION MODELS

The statistical evaluation of multiple-regression models is similar to that discussed in Chapter 4 for simple bivariate regression models. However, some important differences will be brought out in this section. In addition to evaluating the multiple-regression model, we will be comparing these results with a corresponding bivariate model. Thus, in Table 5.2, you see the regression results for both models. The multiple-regression results appear at the bottom of the table.

TABLE 5.2 Regression Results for Multiple and Bivariate Regression Models of New Car Sales (NCS) (c5t2&f4)

The Bivariate Regression Model for New Car Sales

Regression Statistics					
R Square	0.228				
Adjusted *R* Square	0.215				
Standard Error	24,121.582				
Observations	60			DW =	0.283
	Coefficients	Std Error	t Stat	P-value	P/2
Intercept	66,396.257	25,045.655	2.651	0.010	0.005
DPIPC	2.866	0.692	4.144	0.000	0.000
ANOVA	df	SS	MS	F	Sig F
Regression	1	9,991,526,948.715	9,991,526,948.715	17.172	0.000
Residual	58	33,747,342,641.469	581,850,735.198		
Total	59	43,738,869,590.183			

The Multiple Regression Model for New Car Sales

Regression Statistics					
Adjusted *R* Square	0.816				
Standard Error	11,690.644				
Observations	60			DW =	1.376
	Coefficients	Std Error	t Stat	P-value	P/2
Intercept	25,304.350	37,454.530	0.676	0.502	0.251
DPIPC	3.227	0.392	8.240	0.000	0.000
UMICS	1,123.363	210.582	5.335	0.000	0.000
UR	−7,659.605	1,661.586	−4.610	0.000	0.000
PR	−3,216.777	1,412.449	−2.277	0.027	0.013
ANOVA	df	SS	MS	F	Sig F
Regression	4	36,221,956,000.002	9,055,489,000.000	66.257	0.000
Residual	55	7,516,913,590.181	136,671,156.185		
Total	59	43,738,869,590.183			

The First Four Quick Checks in Evaluating Multiple-Regression Models

The first thing you should do in reviewing regression results is to see whether the signs on the coefficients make sense.

As discussed in Chapter 4, the **first** thing you should do in reviewing regression results is to see whether the signs on the coefficients make sense. Let's start with the simple bivariate model, that is:

$$\text{NCS} = b_0 + b_1(\text{DPIPC})$$

We have said that we expect a positive relationship between NCS and disposable personal income. Our expectation is confirmed, since:

$$b_1 = +2.866 > 0$$

The second thing to consider is whether these results are statistically significant at our desired level of confidence.

The **second** thing to consider is whether these results are statistically significant at our desired level of confidence. We will follow the convention of using a 95 percent confidence level and thus a 0.05 significance level. The hypothesis to be tested is as follows:

For DPIPC

$H_0: \beta_1 \leq 0$

$H_1: \beta_1 > 0$

This hypothesis is evaluated using a t-test where the calculated t-ratio is found by dividing the estimated regression coefficient by its standard error (i.e., $t_{calc} = b_i$/s.e. of b_i). The table value of $t(t_T)$ can be found from Table 2.5 at $n - (K + 1)$ degrees of freedom, where n = the number of observations and K = the number of independent variables. For our current problem $n = 60$ and $K = 1$, so df = $60 - (1 + 1) = 58$. We will follow the rule that if df ≥ 30, the infinity row of the t-table will be used. Thus, the table value is 1.645. Note that we have used the 0.05 column, since we have one-tailed tests, and in such cases, the entire significance level (0.05) goes in one tail.

Remember that since the t-distribution is symmetrical, we compare the absolute value of t_{calc} with the table value. For our hypothesis test, the results can be summarized as follows:

For DPIPC

$t_{calc} = 4.144$

$|t_{calc}| > t_T$

$4.144 > 1.645$

$\therefore$ Reject H_0

Because the absolute value of t_{calc} is greater than the table value at $\alpha = 0.05$ and df = 58, we reject the null hypothesis at the 0.05 level for the DPIPC coefficient. Thus, we have statistical support for the notion that there is a positive effect of DPIPC on NCS.

By setting the 95 percent confidence level as our criterion, we are at the same time saying that we are willing to accept a 5 percent chance of error or, alternatively, that we set a 5 percent desired significance level.

The third part of our quick check of regression results involves an evaluation of the coefficient of determination.

The **third** part of our quick check of regression results involves an evaluation of the coefficient of determination, which, you may recall, measures the percentage of the variation in the dependent variable that is explained by the regression model. In Chapter 4, we designated the coefficient of determination as R-squared. If you look at the output in Table 5.2, you will see that R^2 is 0.228. This means

that the bivariate model (or variations in DPIPC) explains 22.8 percent of the variation in NCS.

The **fourth** thing to consider is the Durbin-Watson test for serial correlation. For this model, we see in Table 5.2 that DW = 0.283. From Table 4.11 with n = 60 and k = 1, we find D_l and D_u to be 1.55 and 1.62, respectively. Since our DW of 0.283 < 1.55, we find that this model has positive serial correlation. Thus, the standard errors may be biased downward, making the *t*-ratio larger than it should be. It is then possible that we rejected the null hypothesis in error. This positive serial correlation is obvious in the top graph of Figure 5.4, where you see a series

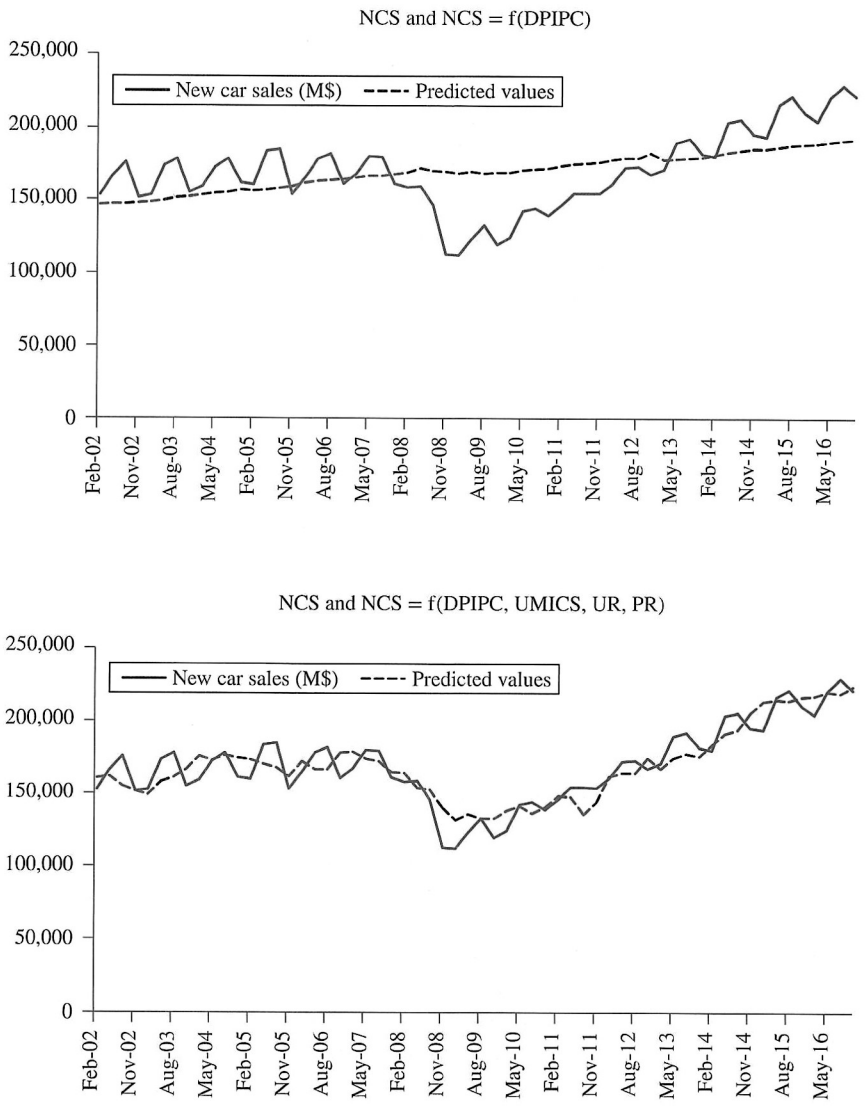


FIGURE 5.4
Two Models for New Car Sales (NCS) (c5t2&f4)

of positive errors followed by a series of negative errors followed by more positive errors. Remember that the errors are calculated as the actual value minus the predicted value for each observation.

Now let us consider the multiple regression model. The process is the same. **First**, are the coefficients logical? That is, do they have the expected signs based on business/economic logic? The results of interest are:

Variable	Coefficient
DPIPC	3.227
UMICS	1,123.363
UR	−7,659.605
PR	−3,216.777

For income (DPIPC), we expect a positive relationship, which we find. The same is true for the University of Michigan Index of Consumer Sentiment (UMICS). For the unemployment rate (UR) and the prime interest rate (PR), we expect and find a negative relationship with sales. So the model is logical. Note that we have said nothing about the intercept (or constant). This is because, for the present purpose, this value has no particular importance for us. It simply places the function at some height in hyperspace and has nothing to do with the causality we are interested in evaluating.

Second, we want to evaluate the statistical significance. As before, we will use a 95 percent confidence level (5 percent significance level). Consider the following values from Table 5.2:

	Coefficients	*t* Stat	P-value	P/2
DPIPC	3.227	8.240	0.000	0.000
UMICS	1,123.363	5.335	0.000	0.000
UR	−7,659.605	−4.610	0.000	0.000
PR	−3,216.777	−2.277	0.027	0.013

Each *t*-statistic is calculated by dividing the coefficient by its standard error. (If you do this by hand, you may get slightly different *t*-statistics due to rounding.) The table *t*-statistic at n-(k + 1) = 60 − (4 + 1) = 55 degrees of freedom is 1.654 using the infinity row of the *t*-table for a one tailed test and a 0.05 significance level. All of our calculated *t*-statistics are larger than 1.645 in absolute terms, so we can conclude that we have statistical support for the hypotheses that DPIPC and the UMICS have a positive relationship with new car sales and that the UR and PR have an inverse (negative) relationship with new car sales.

Let's think about this in a slightly different way: as the *t*-value gets bigger (in absolute value), we move toward a tail of the *t*-distribution, and as we move into the tails of the distribution, the remaining area gets smaller. That remaining area is called a P-value. All statistical software provides a two tailed P-value (although they rarely tell you that explicitly). In business/economics, we often (maybe

usually) have a directional or one-tailed hypothesis such as for DPIPC, UMICS, UR, and PR in our present example. If we divide the two-tailed P-value by two, we get the P-value for a one-tailed test. If P/2 is smaller than our desired significance level, the calculated t-statistic must be larger than the corresponding table value of t. This means that if P/2 is smaller than 0.05 (for our 95 percent confidence level), we can reject the null hypothesis and conclude that the independent variable has a statistically significant influence on the dependent variable. Looking at P-values is what most analysts do. It is quicker and in some cases actually more accurate. In our present example, we see that all of the P/2 values are less than 0.05. Thus, all four independent variables are found to have a significant influence on NCS at a 95 percent level of confidence.

Third, we are interested in the explanatory power of our model. We still want to use the coefficient of determination. But with multiple regression, this is measured by something known as the adjusted R-square. This is a normal part of the output from statistical software. You see in Table 5.2 for the current model the adjusted $R^2 = 0.816$. Thus, this model explains 81.6 percent of the variation in NCS.

In evaluating multiple-regression equations, you should always consider the adjusted R-squared value. The reason for the adjustment is that adding another independent variable will always increase R-squared, even if the variable has no meaningful relation to the dependent variable. Indeed, if we added enough independent variables, we could get very close to an R-squared of 1.00—a perfect fit for the historical period. However, the model would probably work very poorly for values of the independent variables other than those used in estimation. To get around this and to show only meaningful changes in R-squared, an adjustment is made to account for a decrease in the number of degrees of freedom.[3] The adjusted R-squared is often denoted $\bar{R}^2$ (called R-bar-squared or the multiple coefficient of determination).

For our multiple-regression model of new car sales (NCS), we see, in Table 5.2, that the adjusted R-squared is 81.6 percent. Thus, this model explains 81.6 percent of the variation in new car sales. This compares with an R-squared of 22.8 percent for the bivariate model (using only DPIPC as an independent variable).

In looking at regression output, you often see an F-statistic. This statistic can be used to test the following joint hypothesis:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0 \quad \text{or } H_0: R^2 = 0$$

(i.e., all slope terms are simultaneously equal to zero);

$$H_1: \text{All slope terms are not simultaneously equal to zero or } H_1: R^2 \neq 0$$

If the null hypothesis is true, it follows that none of the variation in the dependent variable would be explained by the regression model. It follows that if H_0 is true, the true coefficient of determination would be zero.

[3] These concepts are expanded in J. Scott Armstrong, *Long-Range Forecasting* (New York: John Wiley & Sons, 1978), pp. 323–25, 466.

The F-statistic is calculated as follows:

$$F = \frac{\text{Explained variation}/K}{\text{Unexplained variation}/[n - (K+1)]}$$

The F-test is a test of the overall significance of the estimated multiple regression. To test the hypothesis, this calculated F-statistic is compared with the F-value from Table 5.3 at K degrees of freedom for the numerator and $n - (K + 1)$ degrees of freedom for the denominator.[4] For our current regression, $K = 4$ and $[n - (K + 1)] = 55$, so the table value of F is 2.53 (taking the closest value). In using an F-test, the criterion for rejection of the null hypothesis is that $F_{\text{calc}} > F_T$ (the calculated F must be greater than the table value). In this case, the calculated value is 66.257, so we would reject H_0 (i.e., our equation passes the F-test).

TABLE 5.3 Critical Values of the F-Distribution at a 95 Percent Confidence Level ($\alpha = .05$)

	1*	2	3	4	5	6	7	8	9
1†	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34

(continued on next page)

[4] This F-table corresponds to a 95 percent confidence level ($\alpha = 0.05$). You could use any α value and the corresponding F-distribution.

TABLE 5.3 (continued)

	1*	2	3	4	5	6	7	8	9
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

* Degrees of freedom for the numerator = K
† Degrees of freedom for the denominator = $n - (K + 1)$

Fourth, we want to consider the Durbin-Watson test for serial correlation. In this example, n = 60 and k = 4. Looking at Table 4.11, we find $D_l = 1.44$ and $D_u = 1.73$. Because of our calculated DW = 1.376 (smaller than D_l), we still have positive serial correlation. We will come back to the serial correlation issue again. For now, let us comment that this DW is for a lag of one period. For seasonal data, it is better to use a lag associated with the seasonality: a lag of four for quarterly data and a lag of 12 for monthly data.

The **fifth** step in evaluation for a multiple regression model is new, so we will devote a new topic to that discussion.

Multicollinearity

In multiple-regression analysis, one of the assumptions that are made is that the independent variables are not highly linearly correlated with each other or with linear combinations of other independent variables. If this assumption is violated, a problem known as *multicollinearity* results.

In multiple-regression analysis, one of the assumptions that is made is that the independent variables are not highly linearly correlated with each other or with linear combinations of other independent variables. If this assumption is violated, a problem known as *multicollinearity* results. If your regression results show that one or more independent variables appear not to be statistically significant when theory suggests that they should be, and/or if the signs on coefficients are not logical, multicollinearity may be indicated. Sometimes it is possible to spot the cause of the multicollinearity by looking at a correlation matrix for the independent variables.

To illustrate the multicollinearity problem, suppose that we model new homes sold (NHS) as a function of disposable personal income (DPIPC), the mortgage interest rate (MR), and the gross domestic product (GDP). The model would be:

$$\text{NHS} = b_0 + b_1(\text{DPIPC}) + b_2(\text{GDP}) + b_3(MR)$$

Business and economic logic would tell us to expect a positive sign for b_1, a positive sign for b_2, and a negative sign for b_3. The actual regression results are:

	Coefficient	**t-Ratio**
Constant	1,884	2.09
DPIPC	***-0.01***	-0.21
GDP	0.23	2.83
IR	-147.82	-5.25

We see that the coefficient for DPIPC is negative, which does not make sense. It would be difficult to argue persuasively that NHS would fall as DPIPC rises.

If we look at the correlations between these variables, we can see the source of the problem. The correlations are:

	DPIPC	**GDP**	**IR**
DPIPC	1		
GDP	***0.99***	1	
IR	-0.65	-0.67	1

Clearly, there is a very strong linear association between GDP and DPIPC. In this case, both of these variables are measuring essentially the same thing. There are no firm rules in deciding how strong a correlation is too great. Two rules of thumb, however, provide some guidance. First, we might avoid correlations between independent variables that are close to 1 in absolute value. Second, we might try to avoid situations in which the correlation between independent variables is greater than the correlation of those variables with the dependent variable. One thing to do when multicollinearity exists is to drop all but one of the highly correlated variables. The use of first differences can also help when there is a common trend in the two highly correlated independent variables.

Now let us consider multicollinearity in the context of the NCS multiple regression. We have no reason to suspect a problem since the results are logical and statistically significant as business/economic concepts would suggest. The correlation matrix for the independent variables is shown below.

	UMICS	DPIPC	PR	UR
UMICS	1.00			
DPIPC	-0.16	1.00		
PR	0.25	-0.43	1.00	
UR	-0.68	0.15	-0.62	1.00

Note that no pair of independent variables has a correlation that is close to 1.0. The correlations of 1.00 along the main diagonal are to be expected. They simply tell us that each variable is perfectly correlated with itself. No surprise there. You might ask what does "close to 1.0" mean? There is no concrete answer that

we can use here. It is best not to have any correlation that is in the +/– 0.9s. The +/– 0.8s are also good to avoid, especially if you see more than one in that size range. Ideally, all the correlations would be zero, but that is unrealistic in a business/economic environment. A situation such as shown by the correlation matrix above is good enough.

Serial Correlation: An Extended Look

Serial correlation results when there is a significant time pattern in the error terms of a regression analysis.

The problem known as *serial correlation* was introduced in Chapter 4, where we indicated that serial correlation results when there is a significant time pattern in the error terms of a regression analysis that violates the assumption that the errors are independent over time. Positive serial correlation, as shown in the right-hand graph of Figure 4.11 (page 185), is common with business and economic data.

A test involving comparisons between table values of the Durbin-Watson statistic and the calculated Durbin-Watson statistic is commonly used to detect serial correlation. These comparisons are repeated here, where d_l and d_u represent the lower and upper bounds of the Durbin-Watson statistic from Table 4.12 (page 187).

Value of Calculated Durbin-Watson	*Result*	*Region Designator*
4		
	Negative serial correlation (reject H_0)	A
$4 - d_l$		
	Indeterminate	B
$4 - d_u$		
	No serial correlation (do not reject H_0)	C
2		
d_u		
	Indeterminate	D
d_l		
	Positive serial correlation (reject H_0)	E
0		

In Table 5.2, you see that for the bivariate regression of NCS with DPIPC, the DW is 0.283, indicating positive serial correlation.

For the multiple regression of NCS with DPIPC, UMICS, UR, and PR, the calculated Durbin-Watson statistic is 1.376. (See Table 5.2, which has DW for both the bivariate and the multiple regressions.) This satisfies the region "E" test:

$$\text{DW} < d_l$$
$$1.376 < 1.44$$

where d_l is found from Table 4.11 for $k = 4$ and $N = 60$. Thus, we conclude that this model also has positive serial correlation. This is an improvement in the DW

test for serial correlation, but we have yet to solve the problem. When a regression fails the Durbin-Watson test, the usual interpretation is that this represents the effect of an omitted or unobservable variable (or variables) on the dependent variable. The easiest correction is to collect data on the omitted variable and include it in a new formulation of the model; if the correct variable is added, the serial correlation problem will disappear. However, it is often difficult to identify and measure the missing construct.

In practice, it is often assumed that a first-order check for serial correlation of the residuals will suffice. Remember that the normal Durbin-Watson statistic checks the error terms for serial correlation by comparing errors that are lagged a single period.

When quarterly (or monthly) data are employed, however, the presence of nonsystematic seasonal variation, or an incomplete accounting for seasonality by the included variables, will produce seasonal effects in the error terms, with the consequence that the fourth-order (or 12th order) serial correlation may be significant.

The Durbin-Watson statistic has then been generalized to test for such upper-order instances of serial correlation in the error terms. The fourth-order test statistic has a distribution that differs from that of the normal Durbin-Watson statistic and tables of its critical values as presented in Table 4.7. However, the differences are small, and the user may wish to simply use Table 4.7 to interpret the upper-order Durbin-Watson statistics.[5] When a regression with quarterly (or monthly) data fails the DW(4 or 12) test for serial correlation among the error terms, the usual culprit is that the seasonality in the data has not been fully accounted for by the variables included.

SERIAL CORRELATION AND THE OMITTED-VARIABLE PROBLEM

The most common reason for serial correlation is that an important explanatory variable has been omitted.

The most common reason for serial correlation is that an important explanatory variable has been omitted. To address this situation, it will often be necessary to add an additional explanatory variable to the equation to correct for serial correlation. In Table 5.4 you see that both price and income data are available to use in the model.

In the first regression displayed in Table 5.5, price is used as the single independent variable to explain the firm's sales (the bivariate regression). The results are less than satisfactory on a number of accounts. Most importantly, the sign on the price coefficient is positive, indicating that as price increases, sales also increase. This is inconsistent with business/economic theory and reality. Second, the *R*-squared is quite low, explaining only about 39 percent of the variation in

[5] For a table showing the exact critical values of the Durbin-Watson statistic for quarterly data (both with and without seasonal dummy variables), see the K. F. Wallis article in the Suggested Readings list at the end of this chapter.

TABLE 5.4 Quarterly Data for a Firm's Unit Sales, the Price the Firm Charges for Its Product, and the Income of Potential Purchasers. (c5t4)

PERIOD	Unit SALES	PRICE	INCOME
Mar-13	80	5.00	2620
Jun-13	86	4.87	2733
Sep-13	93	4.86	2898
Dec-13	99	4.79	3056
Mar-14	106	4.79	3271
Jun-14	107	4.87	3479
Sep-14	109	5.01	3736
Dec-14	110	5.31	3868
Mar-15	111	5.55	4016
Jun-15	113	5.72	4152
Sep-15	110	5.74	4336
Dec-15	112	5.59	4477
Mar-16	131	5.50	4619
Jun-16	136	5.48	4764
Sep-16	137	5.47	4802
Dec-16	139	5.49	4916

TABLE 5.5 Statistical Regression Results Based on the Data in Table 5.4. The results shown are from ForecastX™ (c5t4)

The Bivariate Regression

Audit Trail--Coefficient Table (Multiple Regression Selected)

Series Description	Included in Model	Coefficient	Standard Error	*T*-test	P-value	Overall F-test
Sales	Dependent	−51.24	54.32	−0.94	0.36	8.98
Price	Yes	30.92	10.32	3.00	0.01	

Audit Trail--Statistics

Accuracy Measures	Value	Forecast Statistics	Value
AIC	130.02	Durbin Watson (1)	0.34
BIC	130.80		
Mean Absolute Percentage Error (MAPE)	10.67%		
R-Square	39.07%		
Adjusted *R*-Square	39.07%		

Multiple Regression

Audit Trail--Coefficient Table (Multiple Regression Selected)

Series Description	Included in Model	Coefficient	Standard Error	*T*-test	P-value	Overall F-test
Sales	Dependent	123.47	19.40	6.36	0.00	154.86
Price	Yes	−24.84	4.95	−5.02	0.00	
Income	Yes	0.03	0.00	13.55	0.00	

(continued on next page)

TABLE 5.5 (continued)

Audit Trail--Statistics

Accuracy Measures	Value	Forecast Statistics	Value
AIC	86.56	Durbin Watson (1)	1.67
BIC	87.34		
Mean Absolute Percentage Error (MAPE)	2.22%		
R-Square	95.97%		
Adjusted R-Square	95.97%		

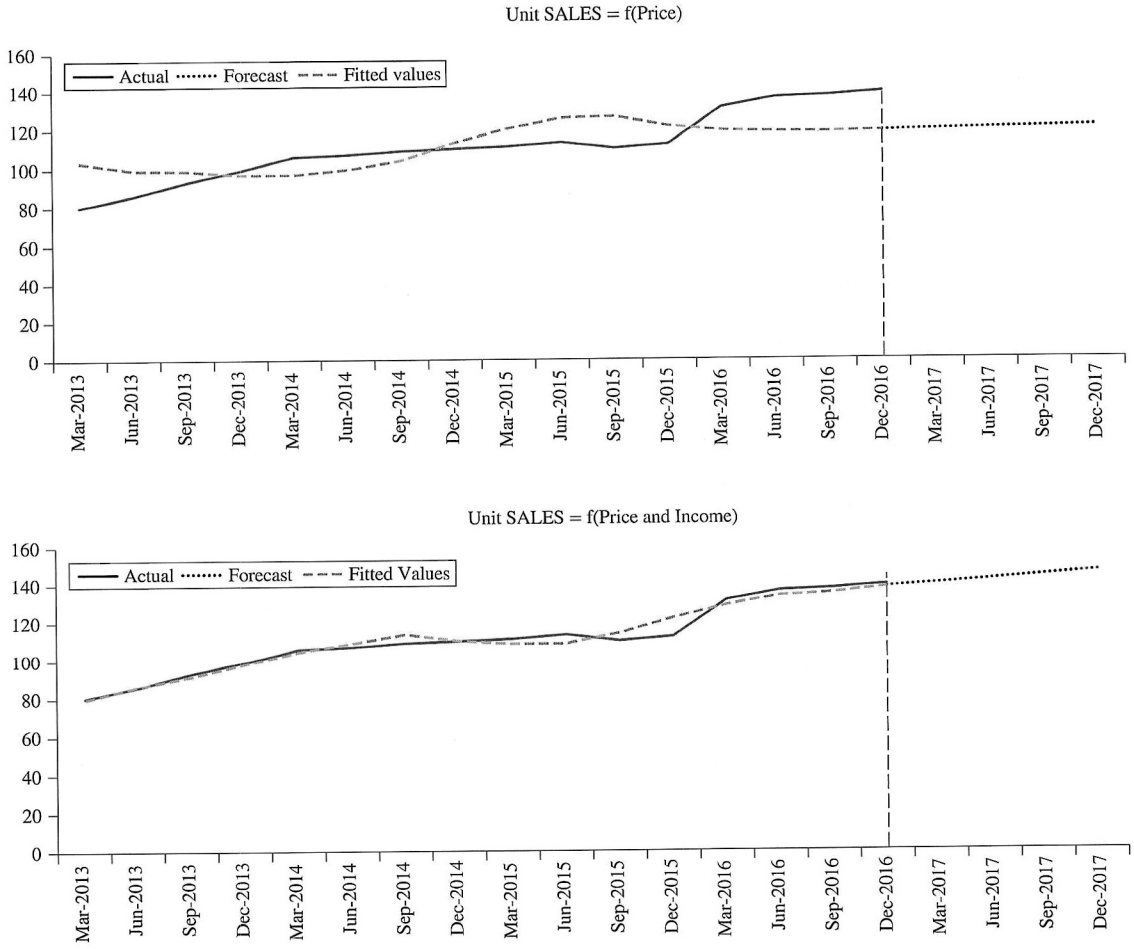
sales. Furthermore, the DW statistic is 0.34, indicating positive serial correlation. See the upper graph in Figure 5.5.

The problem may be that an important variable that could account for the large errors and the incorrect sign of the price coefficient has been omitted from the regression. The second regression in Table 5.5 adds income as a second explanatory variable. The results are dramatic. The adjusted R-squared shows that the model now accounts for about 96 percent of the variation in sales. The signs of both the explanatory variable coefficients are as expected. The price coefficient is negative, indicating that unit sales decrease as price increases, while the income coefficient is positive, indicating that sales of the good rise as incomes increase (which would be reasonable for a "normal" economic good). See the lower graph in Figure 5.5.

The Durbin-Watson statistic (1.67) is within the rule-of-thumb 1.5 to 2.5 range. There does not seem to be serial correlation (and so the R-squared and t-statistics are probably accurate). The formal test for serial correlation requires us to look for the upper and lower values in the Durbin-Watson table (Table 4.11). Note carefully that the appropriate values are 0.95 and 1.54 (i.e., $N = 15$ and column $k = 2$).

Value of Calculated Durbin-Watson	*Result*	*Region Designator and Result*
4		
	Negative serial correlation (reject H_0)	A False
$4 - d_l$		
	Indeterminate	B False
$4 - d_u$		
2	No serial correlation (do not reject H_0)	C True
d_u		
	Indeterminate	D False
d_l		
	Positive serial correlation (reject H_0)	E False
0		

FIGURE 5.5 Graphs for the Two Models from Table 5.5 In the Upper Graph, the Positive Serial Correlation is Clear, Errors are Negative for Early Observations, then Positive for Awhile, then Negative, then Positive. In the Lower Graph, there is Not Such a Distinctive Pattern (c5t4)



Using these Durbin-Watson (DW) values and our calculated value, we see that the following is true: $1.54 < 1.67 < 2$. Thus, we can conclude that no serial correlation is present. Since our result is true for test C, we conclude that no serial correlation is present. Apparently, the addition of the second explanatory variable explained the pattern in the residuals that the Durbin-Watson statistic identified.

Alternative-Variable Selection Criteria

There is a strong tendency for forecasters to use a single criterion for deciding which of several variables ought to be used as independent variables in a regression. The criterion many people use appears to be the coefficient of determination,

or *R*-squared. Recall that *R*-squared is a measure of the proportion of total variance accounted for by the linear influence of the explanatory variables (only *linear* influence is accounted for, since we are using linear least-squares regression). The *R*-squared measure has at least one obvious fault when used in this manner: it can be increased by simply increasing the number of independent variables. Because of this, we proposed the corrected or adjusted *R*-squared, which uses unbiased estimators of the respective variances. Most forecasters use the adjusted *R*-squared to lead them to the correct model by selecting the model that maximizes adjusted *R*-squared. The adjusted *R*-squared measure is based on selecting the correct model by using a quadratic form of the residuals or squared errors in which the true model minimizes those squared errors. But the adjusted *R*-squared measure may not be the most powerful of the measures involving the squared errors.

There are two other model-specification statistics reported by ForecastX™ and other statistical packages that can be of use in selecting the "correct" independent variables.

There are two other model-specification statistics reported by ForecastX™ and other statistical packages that can be of use in selecting the "correct" independent variables. These are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC).[6]

The Akaike information criterion selects the best model by considering the accuracy of the estimation and the "best" approximation to reality. The statistic (which is minimized by the best model) involves both the use of a measure of the accuracy of the estimate *and* a measure of the principle of parsimony (i.e., the concept that fewer independent variables are better than more, all other things being equal). The calculation of the AIC is detailed in Judge and coauthors.[7] We can say that the statistic is constructed so that, as the number of independent variables increases, the AIC has a tendency to increase as well; this means that there is a penalty for "extra" independent variables that must be sufficiently offset by an increase in estimation accuracy to keep the AIC from increasing. In actual practice, a decrease in the AIC as a variable is added indicates that accuracy has increased after adjustment for the rule of parsimony.

In actual practice, a decrease in the AIC as a variable is added indicates that accuracy has increased after adjustment for the rule of parsimony.

The Bayesian criterion is quite similar to the AIC. The BIC uses Bayesian arguments about the prior probability of the true model to suggest the correct model. While the calculation routine for the BIC is quite different from that for the AIC, the results are usually quite consistent.[8] The BIC is also to be minimized, so that, if the BIC decreases after the addition of a new independent variable, the resulting model specification is seen as superior to the prior model specification. Often, AIC and BIC lead to the same model choice.

In a study of the model-selection process, Judge and coauthors created five independent variables that were to be used to estimate a dependent variable. Two of the five independent variables were actually related to the dependent variable,

[6] The Bayesian information criterion is also called the Schwarz information criterion, after its creator.

[7] For a complete description of the calculation routine, see George G. Judge, R. Carter Hill, William E. Griffiths, Helmut Lutkepohl, and Tsoung-Chao Lee, *Introduction to the Theory and Practice of Econometrics*, 2nd ed. (New York: John Wiley & Sons, 1988), chapter 20.

[8] Again see Judge et al. for a complete description of the calculation routine.

while the remaining three were extraneous variables. Various combinations of the five independent variables were used to estimate the dependent variable, and three measures were used to select the "best" model. The three measures used were the adjusted *R*-squared, the AIC, and the BIC.

The correct model containing only the two variables actually related to the dependent variable was chosen 27 percent of the time in repeated experiments by the adjusted *R*-squared criterion. The AIC chose the correct model in 45 percent of the cases, and the BIC chose the correct model in 46 percent of the cases. The results should make the forecaster wary of accepting only the statistical results of what constitutes the best model without some economic interpretation of why a variable is included. It should be clear, however, that the adjusted *R*-squared criterion may not always be the best measure to use in model selection; either the AIC or the BIC would usually be superior. The same study also showed that in 9 percent of the repeated trials the adjusted *R*-squared criterion chose the model with all five variables (i.e., the two "correct" ones and the three extraneous ones). The AIC and the BIC made the same incorrect choice in only 3 percent of the cases.

In Table 5.5, we added a second variable to a regression. When both price and income were included, the AIC decreased from 130.02 to 86.56, and the BIC decreased to 87.34 from 130.80. Apparently, the inclusion of income as a variable was a correct choice.

There is a clear indication of a better identified model if the two competing models differ by more than 10 in their AIC score.

How much of a decrease in the Akaike information criterion constitutes a "better" model? According to Hirotugu Akaike, there is a clear indication of a better identified model if the two competing models differ by more than 10 in their AIC score. If the difference is between 4 and 7, there is much less certainty that a clear winner has emerged. If the difference in AIC scores is 2 or less, then both candidate models have strong support. For this example, the differences in the Akaike scores between the candidate models exceeded 10, and therefore, the second model was clearly the best identified model.

The researcher should not compare the AIC or BIC of one series with the AIC or BIC of another series; the assumption is that models with identical dependent variables are being compared. There is no easy interpretation of the magnitude of the AIC and BIC, nor is one necessary. Only the relative size of the statistics is important.

ACCOUNTING FOR SEASONALITY IN A MULTIPLE-REGRESSION MODEL

Many business and economic data series display pronounced seasonal patterns that recur with some regularity year after year. The pattern may be associated with weather conditions typical of the four seasons of the year. For example, sales of ski equipment would be expected to be greater during the fall and winter (the fourth and first quarters of the calendar year, respectively) than during the spring and summer (the second and third quarters).

Other regular patterns that would be referred to as seasonal patterns may have nothing to do with weather conditions. For example, jewelry sales in the United States tend to be high in November and December because of Christmas shopping and gift giving, and turkey sales are also highest in these months because of traditional Thanksgiving and Christmas dinners.

Patterns such as these are not easily accounted for by the typical causal variables that we use in regression analysis. However, a special type of variable known as a dummy variable can be used effectively to account for seasonality or many other qualitative attributes. The dependent variable in a regression is often influenced not only by continuous variables such as income, price, and advertising expenditures but also by variables that may be qualitative or nominally scaled (such as the season of the year). A dummy variable takes on a value of either 0 or 1. It is 0 if the condition does not exist for an observation, and it is 1 if the condition does exist.

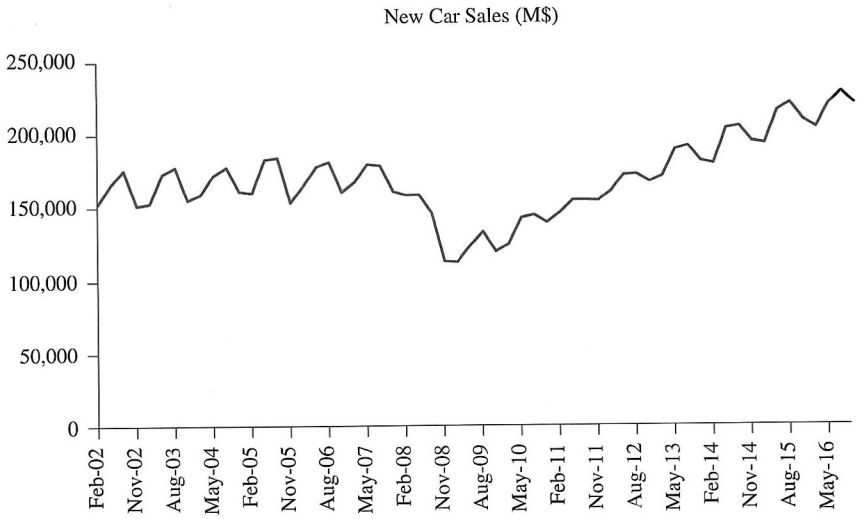
Suppose that we were studying monthly data on turkey sales at grocery stores and we would like to include the November and December seasonality in our model. We could define a dummy variable called M11, for the eleventh month, to be equal to 1 for November observations and 0 otherwise. Another dummy variable, M12, could be defined similarly for December. Thus, for every year these variables would be as follows:

Month	M11	M12	Month	M11	M12
January	0	0	July	0	0
February	0	0	August	0	0
March	0	0	September	0	0
April	0	0	October	0	0
May	0	0	November	1	0
June	0	0	December	0	1

In the regression results, the coefficients for M11 and M12 would reveal the degree of difference in sales for November and December, respectively, compared to other months of the year. In both of these cases, we would expect the coefficients to be positive (indicating that sales in these two months were higher, on average than in the remaining months of the year).

To illustrate very specifically the use of dummy variables to account for and measure seasonality, let us use new cars sold (NCS) in the United States measured in millions of dollars (not seasonally adjusted). These data were first plotted in Figure 5.1 and again here in Figure 5.6. You see in this figure that through the five years, there are lower new car sales during the first quarter than in the other quarters of the year; in most years, there is a peak in sales sometime during the summer months (quarter 3). This pattern is reasonably consistent, although there is variability in the degree of seasonality and some deviation from the overall pattern.

FIGURE 5.6
Total New Cars Sold (NCS) (c5f6)



To account for and measure this seasonality in the NCS regression model, we will use three dummy variables: These will be coded as follows:

Q2 = 1 for quarter 2 and zero otherwise
Q3 = 1 for quarter 3 and zero otherwise
Q4 = 1 for quarter 4 and zero otherwise

Data for new cars sold (NCS), disposable personal income per capita (DPIPC), the University of Michigan Index of Consumer Sentiment (UMICS), the unemployment rate (UR), the bank prime loan rate (PR), and the three seasonal dummy variables are shown in Table 5.6. **Examine the data carefully to verify your understanding of the coding of the seasonal dummy variables.**

Since we have assigned dummy variables for each quarter except quarter 1, the first quarter is the base quarter for our regression model. Any quarter could be used as the base, with dummy variables to adjust for differences attributed to the other quarters. The number of seasonal dummy variables to use depends upon the data. There is one important rule (the Iron Rule of Dummy Variables):

If we have P states of nature, we cannot use more than $P - 1$ dummy variables.

In our current example, $P = 4$, since we have quarterly data, and so we would use only 3 seasonal dummy variables at a maximum. There are 4 states of nature: the 4 quarters in the year. We could use fewer than 3 if we found that all 3 were unnecessary by evaluating their statistical significance by t-tests. But if we violate the rule and use 4 dummy variables to represent all the quarters, we create a situation of perfect multicollinearity (because there is more than one exact relationship among the variables).

TABLE 5.6 **Data for New Cars Sold (NCS), Disposable Personal Income per Capita (DPIPC), the University of Michigan Index of Consumer Sentiment (UMICS), the Unemployment Rate (UR), the Bank Prime Loan Rate (PR), and Three Seasonal Dummy Variables** (c5t6&f7)

Date	New Car Sales (M$)	DPIPC	UMICS	UR	PR	Q2	Q3	Q4
Feb-02	152,641	27,840	93.13	5.70	4.75	0	0	0
May-02	166,036	28,134	94.10	5.83	4.75	1	0	0
Aug-02	175,863	28,168	87.27	5.73	4.75	0	1	0
Nov-02	151,219	28,363	83.83	5.87	4.45	0	0	1
Feb-03	152,843	28,584	79.97	5.87	4.25	0	0	0
May-03	173,360	28,958	89.27	6.13	4.24	1	0	0
Aug-03	178,086	29,539	89.30	6.13	4.00	0	1	0
Nov-03	154,916	29,708	91.97	5.83	4.00	0	0	1
Feb-04	159,086	30,094	98.00	5.70	4.00	0	0	0
May-04	172,174	30,537	93.33	5.60	4.00	1	0	0
Aug-04	178,077	30,800	95.60	5.43	4.42	0	1	0
Nov-04	161,216	31,353	93.87	5.43	4.94	0	0	1
Feb-05	160,070	31,145	94.07	5.30	5.44	0	0	0
May-05	183,594	31,533	90.20	5.10	5.91	1	0	0
Aug-05	184,878	31,961	87.50	4.97	6.43	0	1	0
Nov-05	153,405	32,396	82.43	4.97	6.97	0	0	1
Feb-06	164,864	33,217	88.93	4.73	7.43	0	0	0
May-06	178,363	33,448	83.80	4.63	7.90	1	0	0
Aug-06	181,615	33,696	84.03	4.63	8.25	0	1	0
Nov-06	160,780	33,991	92.47	4.43	8.25	0	0	1
Feb-07	167,528	34,457	92.20	4.50	8.25	0	0	0
May-07	180,006	34,720	86.90	4.50	8.25	1	0	0
Aug-07	179,216	34,918	85.73	4.67	8.18	0	1	0
Nov-07	160,957	35,209	77.50	4.80	7.52	0	0	1
Feb-08	158,369	35,688	72.90	5.00	6.21	0	0	0
May-08	158,761	36,742	59.60	5.33	5.08	1	0	0
Aug-08	146,003	36,175	64.83	6.00	5.00	0	1	0
Nov-08	112,460	35,801	57.67	6.87	4.06	0	0	1
Feb-09	111,819	35,459	58.27	8.27	3.25	0	0	0
May-09	122,950	35,798	68.20	9.30	3.25	1	0	0
Aug-09	132,705	35,546	68.40	9.63	3.25	0	1	0
Nov-09	119,091	35,658	70.17	9.93	3.25	0	0	1
Feb-10	124,117	35,744	73.87	9.83	3.25	1	0	0
May-10	142,278	36,183	73.93	9.63	3.25	0	1	0
Aug-10	144,165	36,397	68.30	9.47	3.25	0	0	1
Nov-10	138,913	36,770	71.27	9.50	3.25	0	0	0
Feb-11	146,009	37,436	73.07	9.03	3.25	1	0	0
May-11	154,335	37,692	71.87	9.07	3.25	0	1	0
Aug-11	154,418	38,017	59.67	9.00	3.25	0	0	1
Nov-11	154,026	38,097	64.80	8.63	3.25	0	0	0
Feb-12	160,743	38,880	75.50	8.27	3.25	1	0	0
May-12	172,000	39,234	76.30	8.20	3.25	0	1	0

(continued on next page)

TABLE 5.6 (continued)

Date	New Car Sales (M$)	DPIPC	UMICS	UR	PR	Q2	Q3	Q4
Aug-12	172,584	39,266	74.97	8.03	3.25	0	0	1
Nov-12	167,223	40,436	79.40	7.80	3.25	0	0	0
Feb-13	171,221	38,828	76.67	7.73	3.25	1	0	0
May-13	189,832	39,010	81.67	7.53	3.25	0	1	0
Aug-13	192,517	39,306	81.57	7.27	3.25	0	0	1
Nov-13	181,804	39,481	76.93	6.93	3.25	0	0	0
Feb-14	179,902	40,049	80.93	6.67	3.25	1	0	0
May-14	203,914	40,693	82.83	6.20	3.25	0	1	0
Aug-14	205,791	41,128	82.97	6.10	3.25	0	0	1
Nov-14	195,535	41,478	89.77	5.70	3.25	0	0	0
Feb-15	194,135	41,447	95.50	5.53	3.25	1	0	0
May-15	216,311	41,966	94.23	5.40	3.25	0	1	0
Aug-15	221,493	42,343	90.73	5.10	3.25	0	0	1
Nov-15	210,181	42,621	91.30	5.00	3.29	0	0	0
Feb-16	204,681	42,807	91.57	4.93	3.50	1	0	0
May-16	221,073	43,265	92.40	4.87	3.50	0	1	0
Aug-16	229,123	43,651	90.33	4.90	3.50	0	0	1
Nov-16	221,392	43,759	93.07	4.70	3.55	1	0	0

Let us now add these dummy variables to the regression model for new cars sold (NCS). Our regression model will include the following independent variables: DPIPC, UMICS, UR, PR, Q2, Q3, and Q4. The model is:

$$\text{NCS} = b_0 + b_1(DPIPC) + b_2(UMICS) + b_3(\text{UR}) + b_4(\text{PR}) + b_5(\text{Q2}) + b_6(\text{Q3}) + b_7(\text{Q4})$$

In this model, we would expect b_1 to have a positive sign (because sales should increase the more disposable income people have), and we would expect b_2 (for the University of Michigan Index of Consumer Sentiment) to have a positive sign because when people feel good about the economy, they are more likely to make a major purchase. We should expect b_3 to have a negative sign (as the unemployment rate rises, fewer people are likely to be in the market for a car). We would expect b_4 to also be negative (as the interest rate rises, cars essentially become more expensive). For b_5 through b_7 (the seasonal dummy variables), we expect the coefficients to be positive since we selected the lowest quarter as the base quarter. It is advisable to pick the lowest season as the base so all other seasonal variables will have positive slopes. This makes evaluating statistical significance much easier.

Regression results for this model are shown in Table 5.7. We see that the model is logical (**evaluation 1**) since all coefficients have the expected signs for their coefficients. All coefficients are statistically significant (**evaluation 2**) at a 95 percent confidence level as indicated by the P/2 values, with the exception of Q2 and Q4, which have P/2 values greater than 0.05. In practice, for dummy variables, we often use a lower confidence level since the variables measure qualitative

TABLE 5.7 Regression Results for New Car Sales (NCS) (c5t7)

Audit Trail – ANOVA Table (Multiple Regression Selected)

Source of variation	SS	df	MS	SEE	Overall F-test
Regression	37,748,211,961.83	7	5,392,601,708.83		46.81
Error	5,990,657,628.35	52	115,204,954.39	10,733.36	
Total	43,738,869,590.18	59			

Audit Trail – Coefficient Table (Multiple Regression Selected)

Series Description	Coefficient	Standard error	*T*-test	P-value	P/2
Intercept	32,398.89	34,589.87	0.94	0.35	0.18
UMICS	1,060.42	195.40	5.43	0.00	0.00
DPIPC	3.14	0.36	8.71	0.00	0.00
UR	−8,232.04	1,535.96	−5.36	0.00	0.00
PR	−3,582.75	1,300.78	−2.75	0.01	0.00
Q2	5,880.09	3,944.63	1.49	0.14	0.07
Q3	14,330.37	4,011.36	3.57	0.00	0.00
Q4	5,138.04	4,006.04	1.28	0.21	0.10

Audit Trail – Statistics

Accuracy Measures	Value	Forecast Statistics	Value
AIC	1,277.42	Durbin-Watson (4)	1.03
BIC	1,279.51	Durbin-Watson (1)	1.16
MAPE	5.05%		
Adjusted R-Square	84.46%		

attributes that are often "fuzzy" constructs. They are not as firm a value as measures such as DPIPC and other quantitative data. The adjusted $R^2 = 84.46$ percent (**evaluation 3**) so over 84 percent of the variation in NCS is explained by the model. Both DW(1) and DW(4) indicate positive serial correlation (**evaluation 4**).

The fifth step in an evaluation is to check for multicollinearity. From the correlations shown below, we see that this is not a problem for this model (**evaluation 5**).

	New Car Sales (M$)	DPIPC	UMICS	UR	PR	Q2	Q3	Q4
New Car Sales (M$)	1.00							
DPIPC	0.48	1.00						
UMICS	0.64	−0.16	1.00					
UR	−0.58	0.15	−0.68	1.00				
PR	−0.02	−0.43	0.25	−0.62	1.00			
Q2	0.04	0.03	0.06	−0.01	−0.03	1.00		
Q3	0.16	0.00	0.03	0.03	0.01	−0.35	1.00	
Q4	−0.07	0.03	−0.11	0.02	−0.01	−0.35	−0.33	1.00

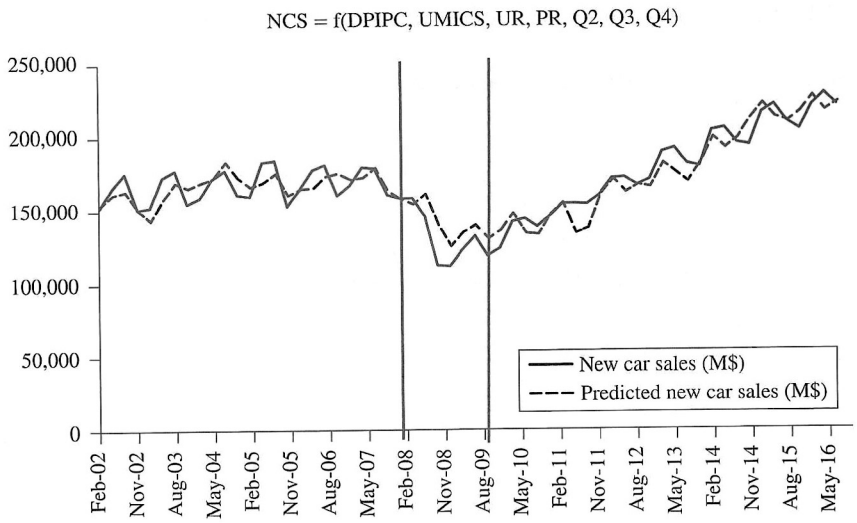


FIGURE 5.7
A Seasonal Model for NCS Actual NCS and the predicted values based on the model shown in the figure title. (c5t6&f7)

The graphic results for this model are shown in Figure 5.7. The distance between the two vertical lines is the time period usually associated with the 2008–2009 recession in the United States. This area also shows a consistent negative error between the actual and predicted value, which may contribute to the serial correlation problem.

The Federal Reserve specification of the 2008–2009 recession is shown in Figure 5.8.

Using a Dummy Variable to Account for a Recession

We see in Figures 5.7 and 5.8 that the 2008–2009 recession seems to correspond with the lower-than-expected NCS in that period. This gives us an opportunity to

FIGURE 5.8 The Federal Reserve Graphic of the 2008–2009 Recession

Source: Federal Reserve Bank of Philadelphia

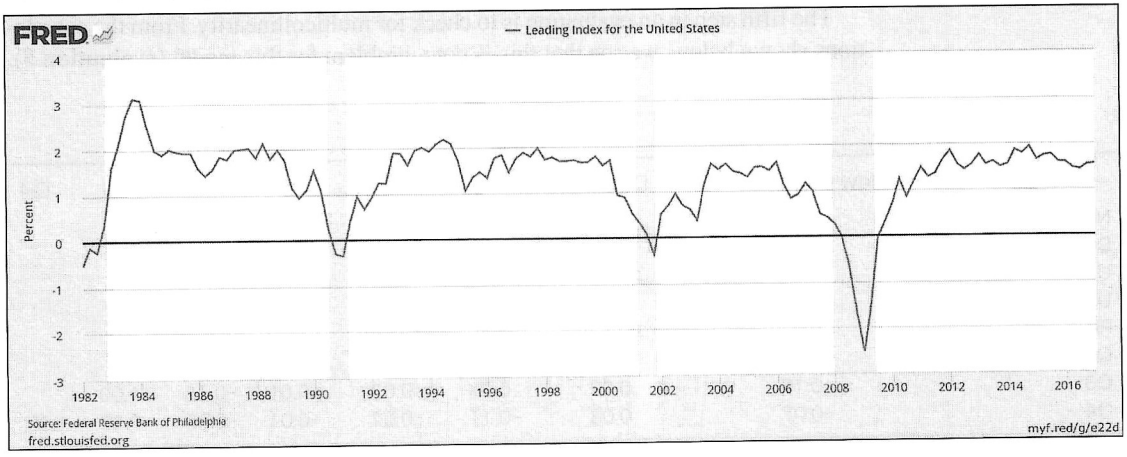


TABLE 5.8 Statistical Results for NCS=f(DPIPC, UMICS, UR, PR, Q2, Q3, Q4, Recession) (c5t8&f9)

Audit Trail – ANOVA Table (Multiple Regression Selected)

Source of variation	SS	df	MS	SEE	Overall F-test
Regression	39,712,666,421.52	8	4,964,083,302.69		62.88
Error	4,026,203,168.66	51	78,945,160.17	8,885.11	
Total	43,738,869,590.18	59			

Audit Trail – Coefficient Table (Multiple Regression Selected)

Series Description	Coefficient	Standard error	*T*-test	P-value	P/2
Intercept	125,763.55	34,208.05	3.68	0.00	0.00
DPIPC	2.82	0.31	9.21	0.00	0.00
UMICS	354.74	214.89	1.65	0.10	0.05
UR	−10,663.56	1,361.70	−7.83	0.00	0.00
PR	−4,705.72	1,100.07	−4.28	0.00	0.00
Q2	6,299.04	3,266.46	1.93	0.06	0.03
Q3	14,896.56	3,322.56	4.48	0.00	0.00
Q4	3,944.55	3,324.83	1.19	0.24	0.12
Recession	−23,253.91	4,661.63	−4.99	0.00	0.00

Accuracy Measures	Value	Forecast Statistics
AIC	1,253.58	Durbin- Watson (4)
BIC	1,255.67	1.00
MAPE	4.21%	Durbin- Watson (1)
Adjusted R-Square	89.35%	1.69

demonstrate another application of a dummy variable. We will create a variable equal to one for all of 2008 and 2009 but zero otherwise. The statistical results are in Table 5.8, and the graphic result is shown in Figure 5.9.

We have already discussed **logic (step 1)** for DPIPC, UMICS, UR, PR, Q2, Q3, and Q4, and see that nothing has changed for this model. All have logical coefficients. For the recession variable, we would expect that other things being equal, NCS would go down during a recession, and thus, we would expect a negative coefficient. Remember that this variable is coded 1 during the recession and zero otherwise. The negative coefficient of -23,253.91 indicates that NCS decreased on average by 23,253.91 million dollars per quarter during the recession.

With regard to **statistical significance (step 2),** there is some question about what one would keep in the model. All the independent variables have low P/2 values (<,0.05) with the exception of UMICS and Q4. Earlier, we discussed a rationale for keeping Q4 in such a model. Regarding the UMICS, the precise P/2 value is 0.052. Some analysts would keep it; others would not. We will keep it.

In terms of **explanatory power (step 3),** we see that the adjusted R^2 indicates that 89.35 percent of the variation in car sales is accounted for by this model. This

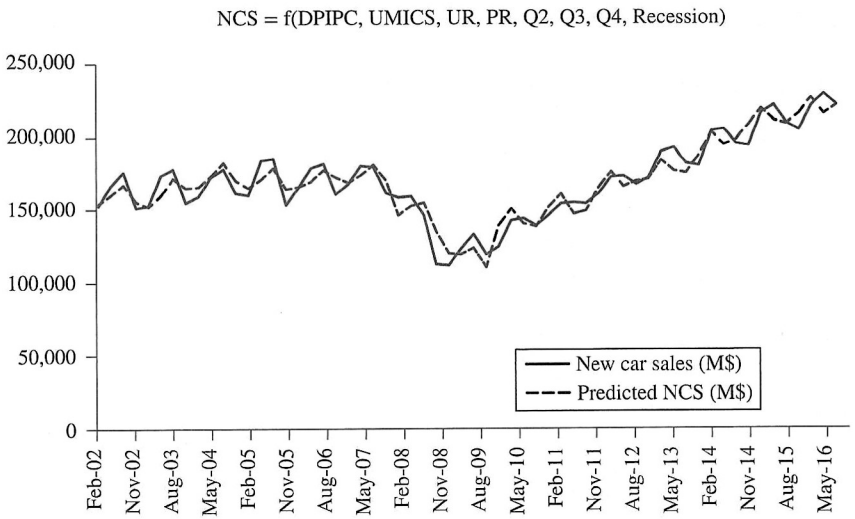


FIGURE 5.9
Actual New Car Sales (NCS) and Regression Model Predictions The predictions are based on the model shown in the figure title. (c5t8&f9)

is higher than the model without the recession variable (without the regression variable, you have seen in Table 5.6 that the adjusted R^2 was 84.46 percent).

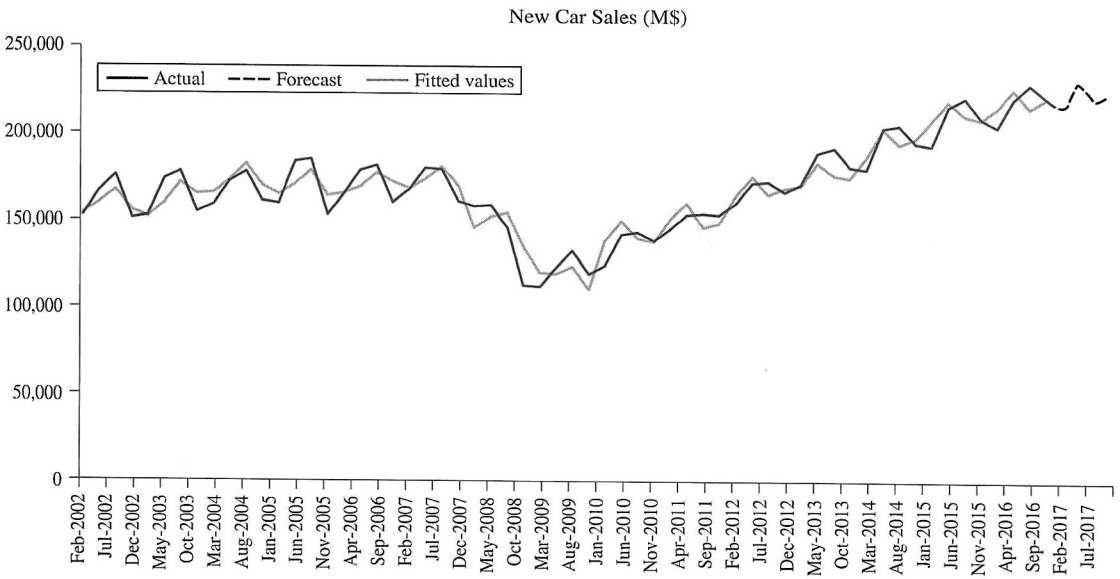
For **serial correlation (step 4),** we see that the DW(4) is 1.00, indicating positive serial correlation. The DW(1) is 1.69, which would indicate an indeterminate conclusion regarding serial correlation. This is a good example of why the seasonal DW is preferable. Given that we have a positive serial correlation, we should consider the consequence. We know that serial correlation causes a downward bias in standard errors, and thus, the *t*-statistics are biased upward. Most of the *t*-statistics are quite high, so this may not be too great a problem. How one thinks about this can be different depending on whether the model is to be used to make strategic decisions regarding the variables or whether one plans to use the model only as a forecasting tool. As long as the model performs well in forecasting, we might be less concerned about some serial correlation. In this case, the model appears to work well. We have a known value for NCS in the first quarter of 2017. It is 212,994 (M$). The percentage error for the first quarter of 2017 is -1.12 percent, which is quite small.

Finally, we will see if there is a **multicollinearity** problem **(step 5)** with adding the recession variable. The correlations between the other independent variables have been shown above, and those are not affected by adding the recession variable. All we need to look at are the correlations of the recession variable with the other variables.

	New Car Sales (M$)	DPIPC	UMICS	UR	PR	Q2	Q3	Q4	Recession
Recession	-0.53	-0.01	-0.62	0.25	-0.07	-0.01	0.00	0.00	1.00

FIGURE 5.10 **The ForecastX™ Multiple Regression Forecast of New Car Sales (M$) for the Four Quarters of 2017** (c5t7&f9)

Source: John Galt Solutions



We see that the recession variable has low correlations with the other independent variables. Thus, we know that we do not have a multicollinearity problem with this model.

The forecast values for 2017 are:

Feb-2017	May-2017	Aug-2017	Nov-2017
215,385.63	231,209.34	221,192.76	224,371.99

The actual, fitted, and forecast values for 2017 are shown in Figure 5.10.

EXTENSIONS OF THE MULTIPLE-REGRESSION MODEL

In some situations, nonlinear terms may be called for as independent variables in a regression analysis. Why? Business or economic logic may suggest that some nonlinearity is expected. A graphic display of the data may be helpful in determining whether the nonlinearity occurs over time. One common cause for nonlinearity is diminishing returns. For example, the effect of advertising on sales may diminish on a dollar-spent basis as increased advertising is used. Another common cause is referred to an Engel's law: As an individual's income doubles, the amount spent on food usually less than doubles (i.e., the proportion spent on food decreases). Both these situations are properly modeled as nonlinearities. In this

TABLE 5.9
Sales Data for 14 Quarters
(c5t9&f11)

Date	Sales
Feb-13	2,010
May-13	1,625
Aug-13	1,612
Nov-13	1,705
Feb-14	1,646
May-14	1,699
Aug-14	1,705
Nov-14	1,795
Feb-15	2,099
May-15	2,294
Aug-15	2,301
Nov-15	2,598
Feb-16	2,689
May-16	2,908

section, we will look at some sales data that are increasing at an increasing rate and compare linear and nonlinear trends. Both of these can be accomplished using ordinary least squares regression. We will use a time index and add the square of a time index as an independent variable to the regression model.

The basic models are:

$$Y = b_0 + b_1(X)$$

and

$$Y = b_0 + b_1(X) + b_2(X^2)$$

where a time index will be the X variable and sales will be the Y variable. So we will have:

$$Sales = b_0 + b_1(Time)$$

and

$$Sales = b_0 + b_1(Time) + b_2(Time^2)$$

The sales data we will use are in Table 5.9 and are graphed in Figure 5.11. As you look at the data in tabular form, it is hard to see whether the relationship over time is linear or nonlinear. However, in the graph, it becomes easy to see that over time, sales are increasing at an increasing rate. A clear nonlinear shape emerges that is not so obvious in the table. This is yet another example of the value of data visualization in a graphic form.

We first estimate the linear trend and second the nonlinear (quadratic) trend. The results of both are shown in Table 5.10 and Figure 5.12.

Look carefully at the statistical results for the two models in Table 5.10. Compare the bold-faced items for the linear regression trend with the corresponding bold-faced items for the nonlinear trend. These comparisons will help you understand the significance of adding a squared term to improve the model. Let's look at all of these comparisons.

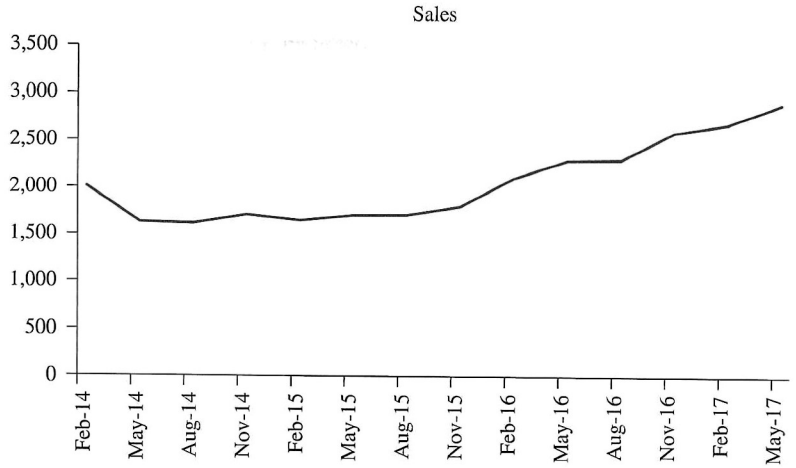


FIGURE 5.11
Fourteen Quarters of Sales Data A visual inspection suggests a nonlinear shape over time. (c5t9&f11)

- First, look at the *t*-tests. You see that in the linear model that the slope for "Time" is very significant with a *t*-value of 5.66 and a P/2 of 0.00. For the nonlinear model, both *t*-tests show high significance with *t*-values for Time and Time^2 of −3.87 and 7.00, respectively.
- Second, look at the coefficients of determination. For the bivariate linear model, we see that *R*-squared is 72.76 percent. For the multiple regression (the quadratic model), the adjusted *R*-squared is 94.09 percent. This represents a big improvement in explanatory power for the nonlinear model.
- Third, look at the MAPEs. For the linear model, the MAPE is 8.64 percent, while for the nonlinear model, the MAPE drops to 4.06 percent. Again we see considerable improvement favoring the nonlinear model.
- Fourth, look at the Durbin-Watson results. For the linear model, the critical values for DW are $d_l = 1.08$ and $D_u = 1.36$. For the linear model, we see that DW = 0.57, which is less than $d_l = 1.08$, indicating **positive serial correlation**. You might have guessed this would be the result by looking at the graph at the top of Figure 5.12. Now, look at the Durbin-Watson statistic for the quadratic model. It is 1.67. The critical values for DW are $d_l = 0.95$ and $D_u = 1.54$. Using the DW test, we have $1.54 < 1.67 < 2.00$, indicating **no serial correlation**. This shows again that the nonlinear (quadratic) model is superior to the linear model.
- Finally, compare the AIC and BIC values. The AIC and BIC statistics for the linear model are 193.01 and 193.66, respectively. For the nonlinear model, the AIC and BIC statistics are 169.27 and 169.91, respectively. For both AIC and BIC, the difference between the linear and nonlinear models is larger than 10, indicating model improvement favoring the nonlinear model.

Clearly, the nonlinear model is superior to the linear model based on all five of the above comparisons. The graphs in Figure 5.12 also support this conclusion. In this example, the graphs make it quite clear which model is superior. However,

TABLE 5.10 **Linear and Nonlinear Regression Trends. The linear trend model is: Sales = f(Time). The nonlinear trend model is: Sales = f(Time, Time^2), which is a quadratic equation** (c5t10f12).

Results for the Model with Time as the Only Independent Variable

Audit Trail—ANOVA Table (Multiple Regression Selected)

Source of variation	SS	df	MS	SEE	Overall F-test
Regression	1,843,110.02	1	1,843,110.02		32.05
Error	690,163.98	12	57,513.67	239.82	
Total	2,533,274.00	13			

Audit Trail — Coefficient Table (Multiple Regression Selected)

Series Description	Coefficient	Standard error	*T*-test	P-value	P/2
Intercept	1,373.93	135.38	10.15	0.00	0.00
Time	90.01	15.90	**5.66**	0.00	**0.00**

Audit Trail—Statistics

Accuracy Measures	Value	Forecast Statistics	Value
AIC	**193.01**	Durbin Watson (1)	**0.57**
BIC	**193.65**		
MAPE	**8.64%**		
***R*-Square**	**72.76%**		

Results for the Model with Time and Time Squared as Independent Variables

Audit Trail—ANOVA Table (Multiple Regression Selected)

Source of variation	SS	df	MS	SEE	Overall F-test
Regression	2,406,605.23	2	1,203,302.62		104.50
Error	126,668.77	11	11,515.34	107.31	
Total	2,533,274.00	13			

Audit Trail—Coefficient Table (Multiple Regression Selected)

Series Description	Coefficient	Standard error	T-test	P-value	P/2
Intercept	1,930.36	99.98	19.31	0.00	0.00
Time	−118.65	30.67	**−3.87**	0.00	**0.00**
Time^2	13.91	1.99	**7.00**	0.00	**0.00**

Audit Trail — Statistics

Accuracy Measures	Value	Forecast Statistics	Value
AIC	**169.27**	Durbin Watson (1)	**1.67**
BIC	**169.91**		
MAPE	**4.06%**		
Adjusted *R*-Square	**94.09%**		

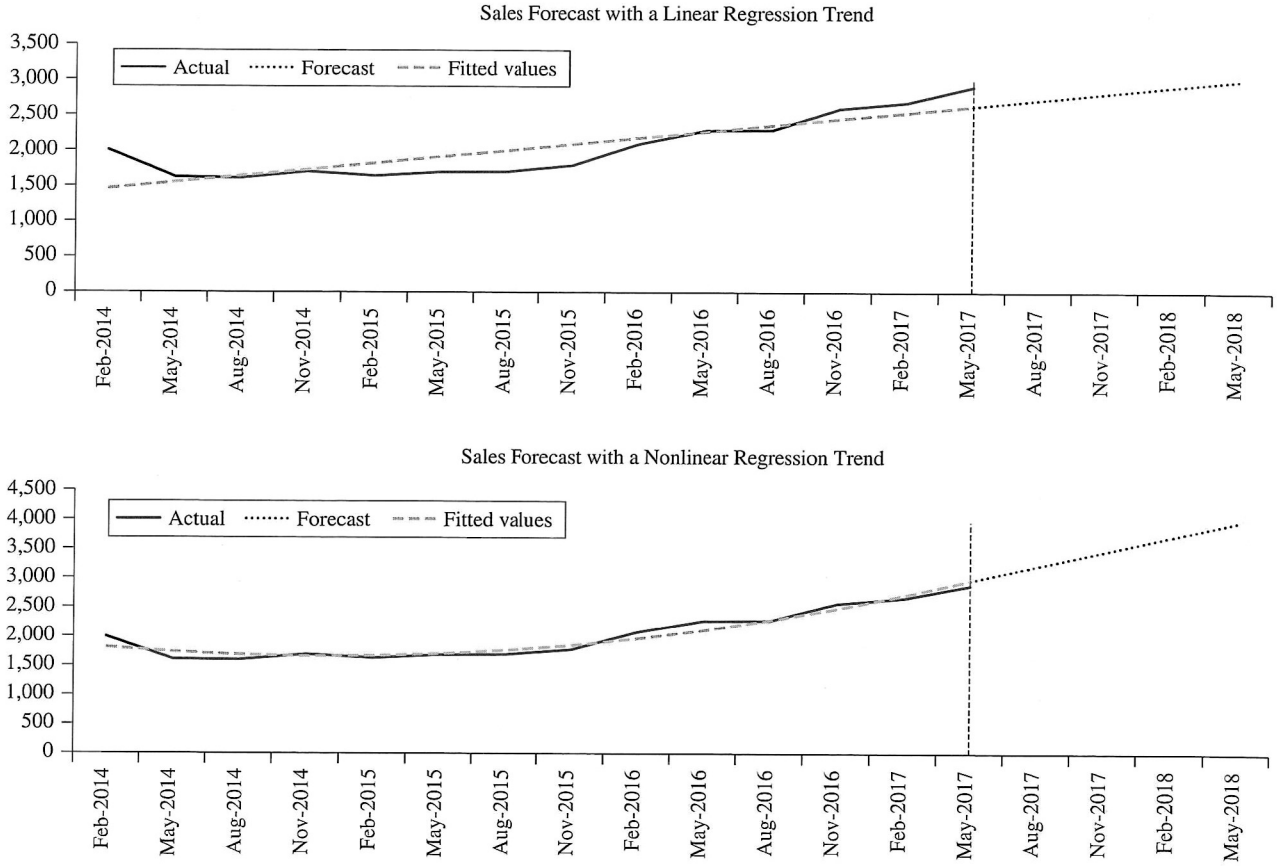


FIGURE 5.12 The Linear And Nonlinear (Quadratic) Forecasts of Sales The linear trend model is: Sales = f(Time). The nonlinear trend model is: Sales = f(Time, Time^2). (c5t10f12).
It appears clear that the nonlinear model is likely to provide the better forecast of sales.

in many cases, the graphs may not make it quite so clear, so it is always good to evaluate the statistical properties of different models.

Using ForecastX™ to generate a forecast using each of the models, we get the forecasts pictured in Figure 5.12. In the top graph, we see that the forecast (dotted line) appears as though it would be too low for all of the four quarters being forecast. On the other hand, in the lower graph, we see that the nonlinear model is likely to give a more reasonable forecast. Of course, we never know for sure until the actual results for the forecast period are known.

ADVICE ON USING MULTIPLE REGRESSION IN FORECASTING

Multiple-regression models are a very important part of the set of tools available to anyone interested in forecasting. Apart from their use in generating forecasts, they have considerable value in helping us to uncover structural relationships between the dependent variable and some set of independent variables. Knowing such relationships helps the forecaster understand the sensitivity of the variable to be forecast to other factors. This enhancement of our understanding of the business environment can only serve to improve our ability to make judgments about the future course of events. It is important not to downplay the role of judgments in forecasting. No one should ever rely solely on some quantitative procedure in developing a forecast. Expert judgments are crucial, and multiple-regression analyses can be helpful in improving your level of expertise.

In developing forecasts with regression models, perhaps the best advice is to follow the "KIS" principle: keep it simple.[9] The more complex the model becomes, the more difficult it is to use. As more causal variables are used, the cost of maintaining the needed database increases in terms of both time and money. Further, complex models are more difficult to communicate to others who may be the actual users of the forecast. They are less likely to trust a model that they do not understand than a simpler model that they do understand.

In evaluating alternative multiple-regression models, is it better to compare adjusted *R*-squared values or mean absolute percentage errors? Remember that *R*-squared relates to the in-sample period, that is, to the past. A model may work well for the in-sample period but not work nearly so well in forecasting. Thus, it is usually best to focus on the MAPE for actual forecasts (note that we say "focus on" and not "use exclusively"). You might track the MAPE for several alternative models for some period to see whether any one model consistently outperforms others in the forecast horizon. Use the AIC and BIC measures to help select appropriate independent variables. It is also desirable periodically to update the regression models to reflect possible changes in the parameter estimates.

[9] This is also called the *principle of parsimony* by Box and Jenkins. G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, 2nd ed. (San Francisco: Holden Day, 1976).

INDEPENDENT VARIABLE SELECTION

Every forecaster who attempts to use demand planning models such as multiple causal regression has difficulty in predicting the turning points in their data with enough accuracy and timeliness to prove useful, and that is the source of most of the errors incurred. While every forecaster subscribes to monitoring certain economic time series, they rarely do it with confidence in the outcome. Finding relevant independent variables with the right periodicity and covering the historic period matching our data is difficult.

The results of the forecast errors we incur are poorly timed promotions, millions lost in safety stocks that are larger than necessary, and loss in market share. How large are the errors forecasters make in the real world? According to the Institute of Business Forecasting (IBF) study[10] by Chaman L. Jain, they are quite large. Jain estimated in 2015 that forecast errors of demand planners at the SKU (stock keeping unit) level average between 27 and 37 percent. For forecasts made at the "category" level, the error averaged between 15 and 26 percent. Not surprisingly, for higher level aggregate forecasts, the errors averaged between 10 and 15 percent. As the level of granularity increased, the average error level also increased. Jain was surprised at the large magnitude of the average errors. Could these errors be reduced?

Many of those errors made by practicing demand planners came from models that had very good fit statistics (recall that fit refers to "in-sample" measurements made with historic data); their MAPEs were low, and their coefficients of multiple determination (*R*-squared) were high. Some of those forecasters may have assumed that a model's fit to the historic data indicated how accurately the model would forecast the future values. So, if the error of the historic fit was 20 percent, then the error of the future forecasts would also be predicted to be about 20 percent. You know that is a very grave error. Remember that the dirty trick of software vendors is to only show potential clients how well the software can fit models to historic data but never show clients how well the software actually forecasts (i.e., predicts in a future period). For that reason, we have suggested using holdouts to form out-of-sample tests. Forecast accuracy will almost always be worse and often much worse than the fit of a model to historic data. That is exactly what the IBF study demonstrates. But what is the reason for the low errors in the historic period and much higher errors in the forecast period?

The primary reason why multiple regression demand planning models fit our historical data so well is that we have complete information available for the historic period on our independent variables; we know for certain the values of the independent right-hand side variables for the entire historic period! However, when we turn the model to forecasting the next 12 to 18 months into the future, the results deteriorate, as demonstrated by the large out-of-sample errors reported in the IBF study.

[10] Chaman L. Jain. *Benchmarking Forecast Errors: Research Report 13* (New York: Institute of Business Forecasting and Planning, 2015).

That means that the indicators we *should* use in our demand planning models had best be leading indicators, those that help us predict in advance what the value we are forecasting might be in a future period. We do not want to choose the forecasting model based solely on the model's "fit to history." It is common practice to choose the model that most closely matches the recent history; we then employ it for creating forecasts of the future. But our objective is not to just get the best fit: Our objective should be to find an appropriate model for forecasting future values. Having perfect fit to historic data is no guarantee that the model will generate good forecasts or is even acceptable for forecasting.

So where do we look for leading indicators, those indicators whose current values help predict future values of the dependent variable? We are helped out these days by firms that provide the service of collecting millions of time series and making them available to data scientists; they provide the raw data we need to estimate models that have a reasonable chance of predicting the future. We call these firms "data consolidators." It's estimated that about 85 percent of a company's performance comes about as a result of factors outside the firm. Internal factors such as price changes, product improvements, and advertising all have an effect on sales and the bottom line, but external factors play the much larger role. Unfortunately, most firms pay a great deal of attention to internal factors when building demand planning models because the data is readily at hand. The external factors that affect the firm are much more difficult to come by; they do not exist in internal databases and documents. By their very nature, they lie outside the firm, and a forecaster is going to have to track down this data; for this reason, they are sometimes ignored.

This is where the role of the data consolidator comes in. What types of external information provided by a data consolidator might be useful?

Think of the firm's inside information as the small slice of the circle in Figure 5.13 (say, 15 percent of the relevant information); that is the information

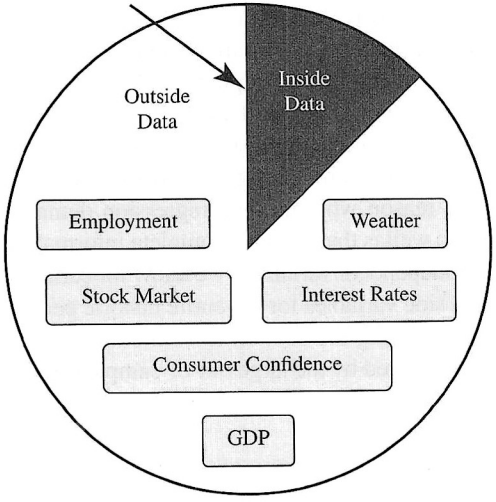


FIGURE 5.13
Only About 15 Percent of the Relevant Information to use in Constructing a Forecast is Data Internal to the Company

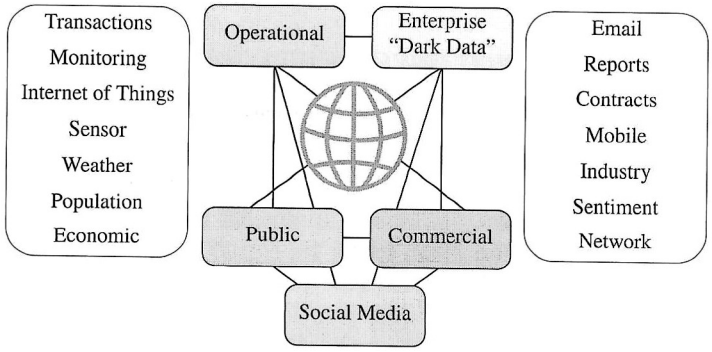


FIGURE 5.14
There are Many Sources and Types of Outside Data

we have internally that we know will help us build the forecast. But the remainder of the circle includes the 85 percent of the information that would be useful to us but to which we do not have convenient access.

Where is this outside data? As Figure 5.14 shows, it comes from many sources: government statistics that are public, weather data, economic statistics from both public and private sources, data from industry organizations, and, increasingly today, data from sensors, the Internet of Things, and social media posts. We not only need information on outside data, we also need information on outside data that are specifically leading indicators for the particular item we are forecasting. The example in the next section might make this clear.

THE GAP EXAMPLE (WITH LEADING INDICATORS)

Firms have always had access to some outside data. The U.S. Bureau of Economic Analysis and the U.S. Census Bureau provide many useful time series, albeit not in easily searchable and ready-to-use formats. By using a firm such as Prevedere (http://www.prevedere.com/), we can open a window to over 3.5 million time series worldwide. Some of these time series are provided by governments, some are proprietary and have been purchased by Prevdere, and others are available from industry groups. The first feature of the service provided by firms such as Prevedere is to collect and, more importantly, organize these time series so that they become searchable and useful for selection as independent variables (i.e., leading indicators).

But how could we possibly select from 3.5 million series? Even scrolling through the names of the different series would be time-consuming and haphazard. We need a tool that will aid us in selecting only those indicator series that will be helpful in our particular situation; we need indicators that "lead" our target series and indicators that are highly predictive (i.e., highly related to the target series). The advantage in using a tool such as Prevedere is that in addition to the data, Prevedere has built into it a search mechanism that helps build the model. The indicators (all 3.5 million) are arranged in tags (Figure 5.15).

FIGURE 5.15
A Portion of the Tags that Allow Users to Narrow the Selection of Useful Indicator Variables

Source: Prevedere Software, Inc.

Some of the tags relate to geographic areas of the world that might focus your search, while others represent industries or product types that could be appropriate. If we were to drill down on "consumer goods" as a category of interest, that would narrow down the search to 21,936 indicator time series (Figure 5.16).

But even searching over 21,936 indicators would be a problem. The second feature of the Prevedere type service is that the program contains a search mechanism that will search over the range of "areas" (such as consumer goods) and pull out indicators that match the pattern exhibited by your data. Even more importantly, the package matches the pattern in your data (the example used here is the Gap data) with calculated best lead times. Recall that to actually "forecast," you will need to have a current indicator value that predicts a future target value; if the variables have lead times, they can be used at least for the value of the lead time into the future (say, three quarters). The optimal lead times for the different indicators are calculated and displayed in the Prevedere engine (Figure 5.17).

FIGURE 5.16
A Few of the "Consumer Goods" Indicators from the List of 21,936 Indicators in this Single Tag

Source: Prevedere Software, Inc.

Page 1 Of 878 - 21,936 Total Matches For "#ConsumerGoods"

G **Goods, Value Of Imports For Euro Area**
International Monetary Fund

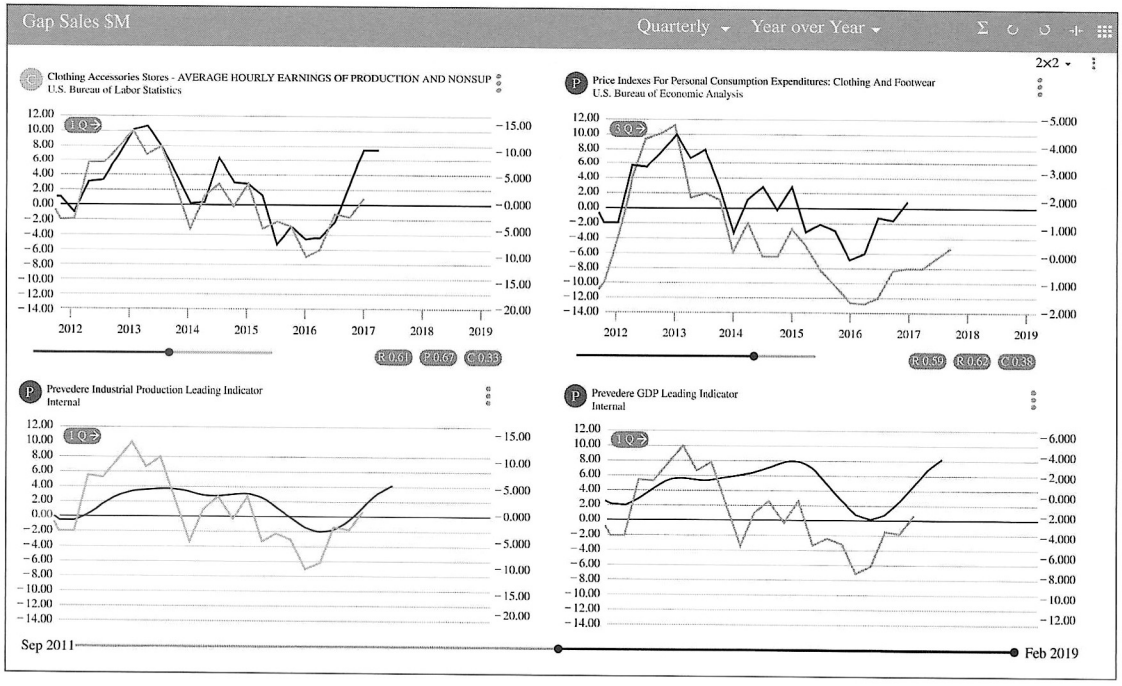
E **Euro Area (18 Countries): Production In Industry: Volume Index Of Production; MIG - Consumer Goods (Except Food, Beverages And Tobacco)**
Eurostat

E **European Union (15 Countries): Production In Industry: Volume Index Of Production; MIG - Consumer Goods**
Eurostat

E **European Union (15 Countries): Production In Industry: Volume Index Of Production; MIG - Consumer Goods (Except Food, Beverages And Tobacco)**
Eurostat

FIGURE 5.17 **Some of the Indicators, with Lead Times Shown in Quarters, for the Gap Data**

Source: Prevedere Software, Inc.



One of the selected indicators is "Price Indexes for Personal Consumption Expenditures," which is obtained from the U.S. Bureau of Economic Analysis. That series was chosen by the Prevedere software because it has a pattern that, with a lead time of three quarters, closely matches our target variable: Gap sales. This variable with an offset of three quarters is then used in the forecast model reported. Notice that all the indicators chosen (although we only display four here) do not have the same offset as the price indexes series; it is not unusual that different time series will have a distinctly different relationship to Gap sales. The regression model that we would build for Gap sales should then be "segmented." In other words, we would use all the indicators with at least a one-quarter lead to forecast for the first future quarter; we would then use all the indicators with at least a two-quarter lead to forecast out for the second future quarter; and so on. Each "segment" is actually a separate multiple regression just like those described earlier in this chapter. As we move from the first segment to the second segment, we would "lose" the indicators that have only a one-quarter lead; we cannot use those indicators to forecast for the second quarter in the future. The segments are constructed using a rolling holdout procedure throughout the historic period to ensure that the model was accurate (as opposed to simply "fitting" well).

Figures 5.18 and 5.19 display the first two segments of a Gap model. The first segment used five indicators; the second segment, however, used only four

FIGURE 5.18 **Segment 1 of the Gap Forecasting Model Produced in Prevedere**

Source: Prevedere Software, Inc.

Statistics

SEGMENT 1 SEGMENT 2 SEGMENT 3

Segment For 04/2017 to 04/2017

P-Value 0.000 F Statistic 34.579 R-Squared 0.836 Adjusted R-Squared 0.812 Predictive R-Squared 0.773 Est. Of Standard Error 2.189

Value	Standard Error	T Value	90% Conf. Interval	95`% Conf. Interval
Intercept				
−1.380	0.840	−1.644	[−2.7619, .0010]	[−3.0266, .2657]
Clothing accessories stores - AVERAGE HOURLY EARNINGS OF PRODUCTION AND NONSUPERVISORY EMPLOYEES				
0.249	0.079	3.147	[.1190, .3796]	[.0940, .4046]
Architectural Billings Index - New Projects Inquiries				
0.077	0.044	1.775	[.0057, .1491]	[−.0080, .1629]
Price Indexes for Personal Consumption Expenditures: Clothing and footwear				
1.104	0.299	3.692	[.6121, 1.5955]	[.5179, 1.6897]
Corporate profits with inventory valuation adjustments: Domestic industries: Nonfinancial: Retail trade				
0.056	0.026	2.139	[.0129, .0987]	[.0047, .1069]
Industrial Capacity - Apparel				
−0.035	0.072	−0.484	[−.1524, .0832]	[−.1750, .1057]

FIGURE 5.19 Segment 2 of the Gap Forecasting Model Produced in Prevedere Note that this is a different regression from the one used in segment 1.

Source: Prevedere Software, Inc.

Statistics

SEGMENT 1 SEGMENT 2 SEGMENT 3

Segment For 07/2017 to 10/2017

P-Value 0.000 F Statistic 32.488 R-Squared 0.788 Adjusted R-Squared 0.764 Predictive R-Squared 0.734 Est. Of Standard Error 2.452

Value	Standard Error	T Value	90% Conf. Interval	95% Conf. Interval
Intercept				
0.237	0.744	0.319	[-.9860, 1.4608]	[-1.2203, 1.6951]
Architectural Billings Index - New Projects Inquiries				
0.091	0.049	1.870	[.0110, .1708]	[-.0044, .1861]
Price Indexes for Personal Consumption Expenditures: Clothing and footwear				
1.629	0.278	5.868	[1.1727, 2.0861]	[1.0852, 2.1736]
Corporate profits with inventory valuation adjustments: Domestic industries: Nonfinancial: Retail trade				
0.046	0.029	1.570	[-.0022, .0932]	[-.0113, .1023]
Industrial Capacity - Apparel				
0.118	0.059	1.997	[.0208, .2150]	[.0022, .2336]

indicators. "Clothing Accessories Stores" was the indicator that was dropped when segment two was estimated; this indicator had only a one-quarter lead on the target variable (i.e., Gap sales). Using just a very simple model like the one described here, how well does the model perform?

The MAPE values shown in Figure 5.20 for the model are quite low, in the 1 percent to 2 percent range. The small number of indicators used is able to accurately match the actual Gap sales pattern. RaceTrac Petroleum used the same technique, identifying outside indicators that matched the pattern of their sales with some lead time, to significantly improve their forecasting effort. According to Brad Galland, RaceTrac director of financial planning, "We felt we were at the mercy of the market and specifically looked for a way to get our hands on external data—the right external data, though. We had internal sales data from the vendors, but this did not help us accurately project sales going forward. As a result, we were operating the business with hindsight as the guide. Ultimately, we wanted a simple model so we could know which product categories in different regions would be affected by different economic and other factors." That model for RaceTrac, using the method shown here, included economic data, weather patterns, customer demographics, competitor moves, and even social and cultural trend indicators.

The forecast for our particular example, as well as the actual Gap sales for 2016, are shown in Figure 5.21.

FIGURE 5.20 Gap Model Performance MAPE values are shown in this plot.

Source: Prevedere Software, Inc.

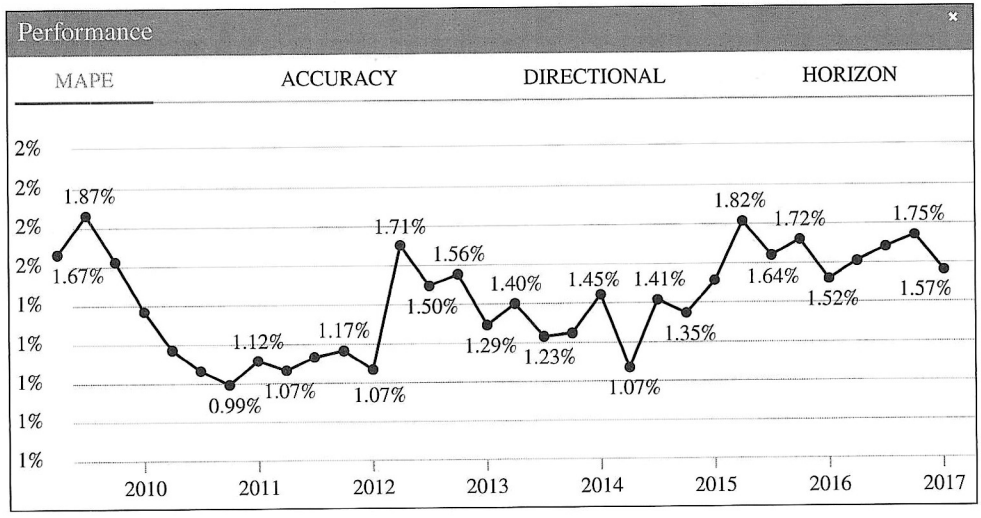
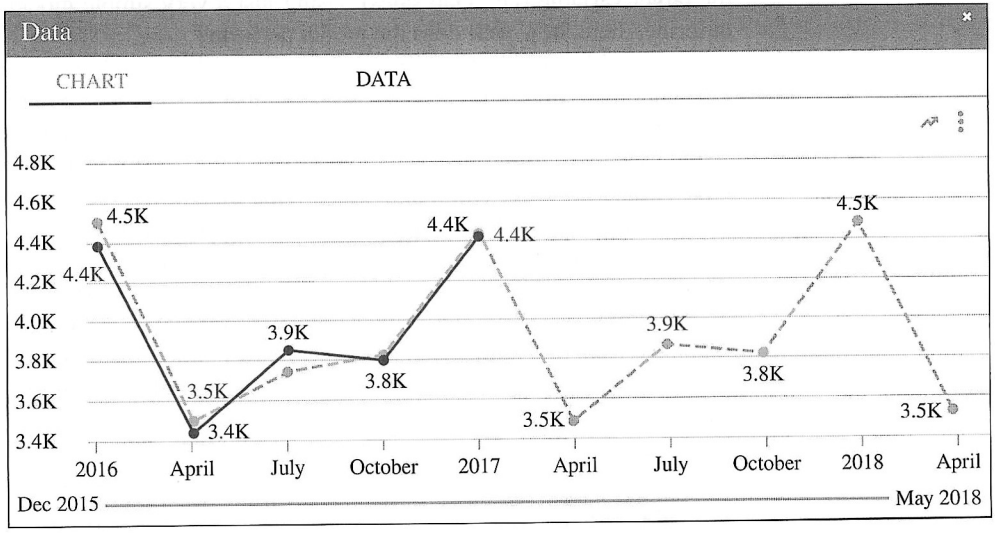


FIGURE 5.21 The Forecast: Gap Actual Sales (solid line) and Forecasted Sales (Dotted Line)

Source: Prevedere Software, Inc.



The key point to remember in constructing demand planning models with multiple regression is to realize that a great deal of the explanatory power of a good regression forecasting model will come from indicators that lie outside the firm's own data. The failure to recognize the reality that a firm exists in a particular economy in a particular location and at a particular moment in time will probably result in producing models much less capable than they could be. Finding those outside indicators has been difficult in the past, but with companies such as Prevedere[11] providing the data and a selection tool, the job is much less difficult today.

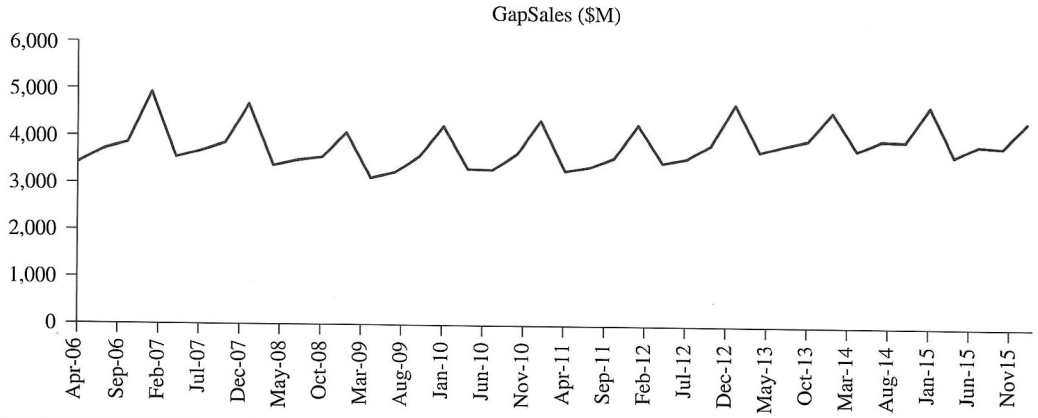
Integrative Case

The Gap

FORECASTING THE GAP SALES DATA WITH A MULTIPLE-REGRESSION MODEL

The sales of The Gap stores in millions of dollars for the 44 quarters covering their Q1 of 2006 (noted in the data file as April 2017) through their Q4 or 2016 (noted in the data file as January 2017) are shown in the graph below. The Gap fiscal year ends in January, and the new fiscal year begins in February. Recall that The Gap sales data are quite seasonal.

(c5Gap)



[11] There are other companies that provide services somewhat like Prevedere. To see a list, search "Prevedere competitors."

Case Questions

1. Have The Gap sales generally followed a linear path over time? Does the graph suggest to you that some accommodation for seasonality should be used in any forecast?
2. Based on the data provided, develop a multiple regression of model for nonseasonally adjusted Gap sales as the basis to forecast sales for 2017.
3. Discuss the MAPE for the historical period (2006 through 2015 fiscal years).
4. Does the model have a serial correlation problem?

Solutions to Case Questions

1. The Gap sales appear to have followed a highly seasonal pattern over time. The peak season appears to be consistent during their fourth quarter (November, December, and January). This would be due to holiday buying by consumers. The data show that Gap sales have been relatively flat during this time frame, but with seasonality.
2. The raw (or nonseasonally adjusted) The Gap sales were used as a dependent variable in a multiple regression that includes the following explanatory (or independent) variables:

$DPIPC$ = Disposable personal income per capita
$DPIPC^2$ = Disposable personal income per capita squared (to help with serial correlation)
Q2 = A seasonal dummy variable for quarter 2
Q3 = A seasonal dummy variable for quarter 3
Q4 = A seasonal dummy variable for quarter 4
UMICS = University of Michigan Index of Consumer Sentiment

The regression results follow:

Audit Trail—ANOVA Table (Multiple Regression Selected)

Source of variation	SS	df	MS	SEE	Overall F-test
Regression	7,192,813.39	6	1,198,802.23		50.49
Error	783,459.99	33	23,741.21	154.08	
Total	7,976,273.38	39			

Audit Trail—Coefficient Table (Multiple Regression Selected)

Series Description	Coefficient	Standard error	T-test	P-value	P/2
Intercept	−11,604.05	8,036.28	−1.44	0.16	0.08
DPIPC	0.72	0.42	1.72	0.10	0.0475
UMICS	18.83	3.80	4.95	0.00	0.00
Q2	170.27	69.52	2.45	0.02	0.01
Q3	315.34	69.50	4.54	0.00	0.00
Q4	1,002.77	69.53	14.42	0.00	0.00
DPIPC2	−0.000009	0.000006	−1.69	0.10	0.0498

Audit Trail—Statistics

Accuracy Measures	Value
AIC	510.82
BIC	512.51
MAPE	2.95%
Adjusted R-Square	88.39%

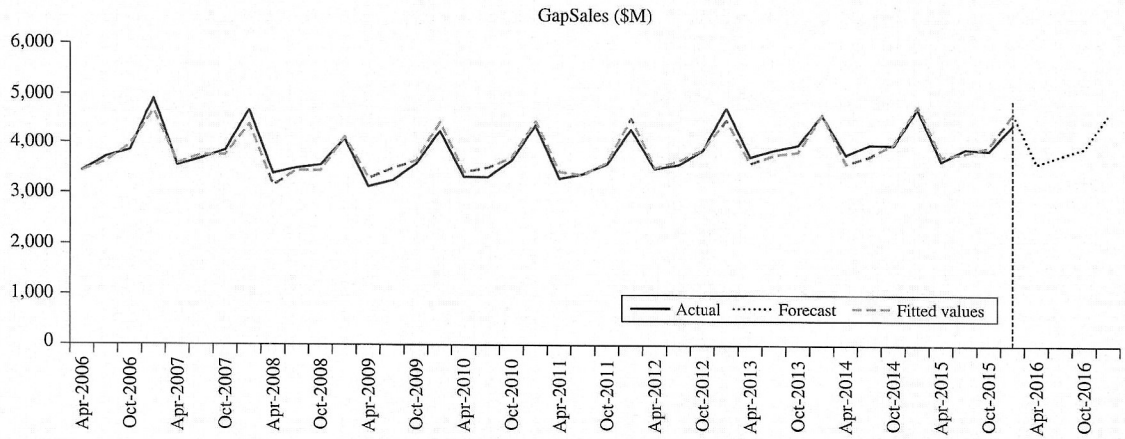
Forecast Statistics	Value
Durbin Watson (4)	1.57

All of the independent variables are logical and statistically significant at our usual 95 percent confidence level. Note that the squared term for DPIPC has a negative coefficient. This suggests diminishing returns to income. The adjusted *R*-squared tells us that this model explains about 88.49 percent of the variation in Gap sales over this time period. The DW(4) value of 1.51 falls between $D_l = 1.29$ and $D_u = 1.78$, so the result is indeterminate with respect to serial correlation. However, since all the P/2 values are very small, it is unlikely that there is a meaningful upward bias to the *t*-test values. The following correlation matrix shows that there is not a multicollinearity problem:

	GapSales ($M)	DPIPC	UMICS	Q2	Q3	Q4	DPIPC^2
GapSales ($M)	1.00						
DPIPC	0.26	1.00					
UMICS	0.41	0.36	1.00				
Q2	−0.29	−0.01	−0.08	1.00			
Q3	−0.08	0.02	−0.02	−0.33	1.00		
Q4	0.85	0.08	0.08	−0.33	−0.33	1.00	
$DPIPC^2$	0.27	1.00	0.38	−0.01	0.02	0.08	1.00

The high correlation between DPIPC and $DPIPC^2$ is to be expected. These two variables are not measuring different constructs but rather measure income in a quadratic form.

A graph of the Gap sales with multiple regression results follows:



(c5Gap)

3. The MAPE is shown in the ForecastX™ results above. The exact calculations are shown below:

Dates	Actual Sales Data	Fitted Data	Error = (A-F)	Absolute Error	Absolute % Error
Apr-2006	3,441.00	3,432.73	8.27	8.273	0.240
Jul-2006	3,714.00	3,627.45	86.55	86.552	2.330
Oct-2006	3,851.00	3,951.77	-100.77	100.769	2.617
Jan-2007	4,919.00	4,659.00	260.00	259.998	5.286
Apr-2007	3,549.00	3,590.33	-41.33	41.334	1.165
Jul-2007	3,685.00	3,755.33	-70.33	70.327	1.908
Oct-2007	3,854.00	3,758.19	95.81	95.811	2.486
Jan-2008	4,675.00	4,375.56	299.44	299.442	6.405
Apr-2008	3,384.00	3,146.14	237.86	237.859	7.029
Jul-2008	3,499.00	3,451.37	47.63	47.633	1.361
Oct-2008	3,561.00	3,443.55	117.45	117.446	3.298
Jan-2009	4,082.00	4,127.53	-45.53	45.527	1.115
Apr-2009	3,127.00	3,297.24	-170.24	170.241	5.444
Jul-2009	3,245.00	3,486.94	-241.94	241.944	7.456
Oct-2009	3,589.00	3,652.59	-63.59	63.587	1.772
Jan-2010	4,236.00	4,414.91	-178.91	178.910	4.224
Apr-2010	3,329.00	3,417.88	-88.88	88.880	2.670
Jul-2010	3,317.00	3,500.23	-183.23	183.227	5.524
Oct-2010	3,654.00	3,707.12	-53.12	53.124	1.454
Jan-2011	4,364.00	4,438.96	-74.96	74.958	1.718
Apr-2011	3,295.00	3,425.84	-130.84	130.845	3.971
Jul-2011	3,386.00	3,368.89	17.11	17.109	0.505
Oct-2011	3,585.00	3,613.24	-28.24	28.242	0.788
Jan-2012	4,283.00	4,502.16	-219.16	219.160	5.117
Apr-2012	3,487.00	3,508.38	-21.38	21.381	0.613
Jul-2012	3,575.00	3,645.75	-70.75	70.747	1.979
Oct-2012	3,864.00	3,874.84	-10.84	10.841	0.281
Jan-2013	4,725.00	4,470.39	254.61	254.610	5.389
Apr-2013	3,729.00	3,608.93	120.07	120.067	3.220
Jul-2013	3,868.00	3,774.43	93.57	93.572	2.419
Oct-2013	3,976.00	3,826.86	149.14	149.141	3.751
Jan-2014	4,575.00	4,585.25	-10.25	10.251	0.224
Apr-2014	3,774.00	3,600.17	173.83	173.833	4.606
Jul-2014	3,981.00	3,744.44	236.56	236.556	5.942
Oct-2014	3,972.00	3,994.29	-22.29	22.294	0.561
Jan-2015	4,708.00	4,769.62	-61.62	61.618	1.309
Apr-2015	3,657.00	3,744.35	-87.35	87.352	2.389
Jul-2015	3,898.00	3,813.18	84.82	84.823	2.176
Oct-2015	3,857.00	3,940.54	-83.54	83.542	2.166
Jan-2016	4,385.00	4,608.63	-223.63	223.625	5.100
				MAPE =	2.950

The MAPE is 2.95 percent. This is relatively low and a good metric to track over time.

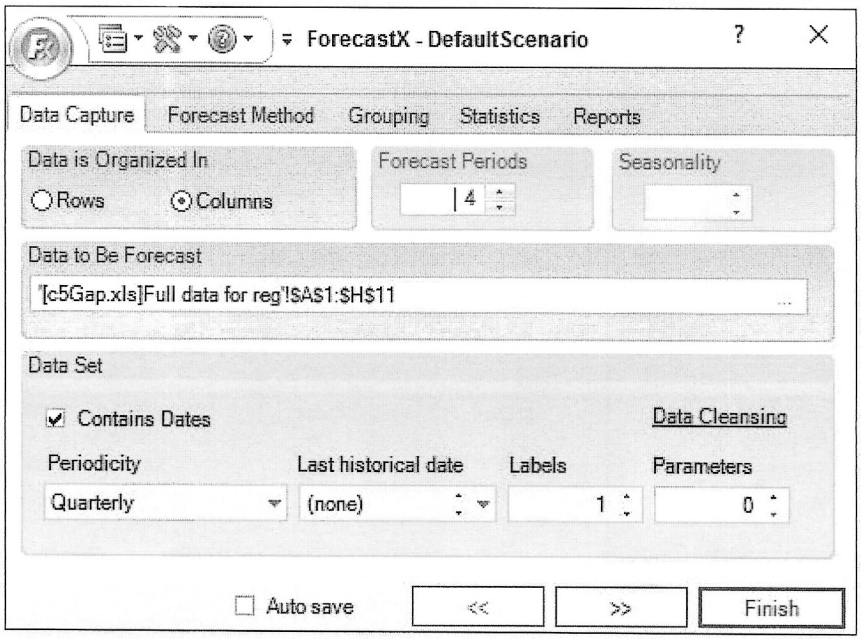
4. The DW(4) is 1.57, which falls in the indeterminate range. Without the squared term for DPIPC (DPIPC$^{2)}$, the DW(4) would be 1.39, indicating positive serial correlation.

USING FORECASTX™ TO MAKE MULTIPLE-REGRESSION FORECASTS

As usual, begin by opening your data file in Excel and start ForecastX™. Place your cursor in a cell with the data to be forecast (cell B6 in this example).

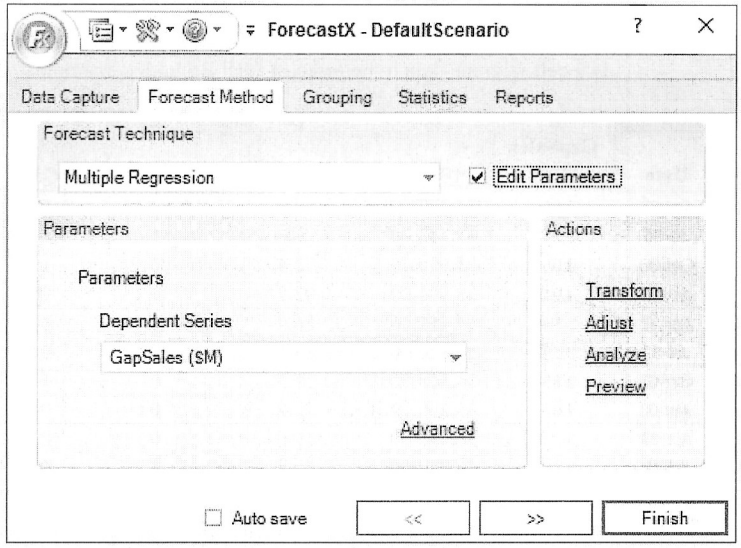
	A	B	C	D	E	F	G	H
1	**Date**	**GapSales ($M)**	**DPIPC**	**UMICS**	**Q2**	**Q3**	**Q4**	**DPIPC^2**
2	Apr-06	3441	33,217.00	83.8	0	0	0	1,103,369,089
3	Jul-06	3,714	33,448.00	84	1	0	0	1,118,768,704
4	Oct-06	3,851	33,696.00	92.4	0	1	0	1,135,420,416
5	Jan-07	4,919	33,991.00	92.2	0	0	1	1,155,388,081
6	Apr-07	3,549	34,457.00	86.9	0	0	0	1,187,284,849
7	Jul-07	3,685	34,720.00	85.7	1	0	0	1,205,478,400
8	Oct-07	3,854	34,918.00	77.5	0	1	0	1,219,266,724
9	Jan-08	4,675	35,209.00	72.9	0	0	1	1,239,673,681
10	Apr-08	3,384	35,688.00	59.6	0	0	0	1,273,633,344
11	Jul-08	3,499	36,742.00	64.8	1	0	0	1,349,974,564

In the **Data Capture** dialog box, verify the data you want to use, as shown below. Check that the periodicity and other features are what you want. Then click the **Forecast Method** tab.



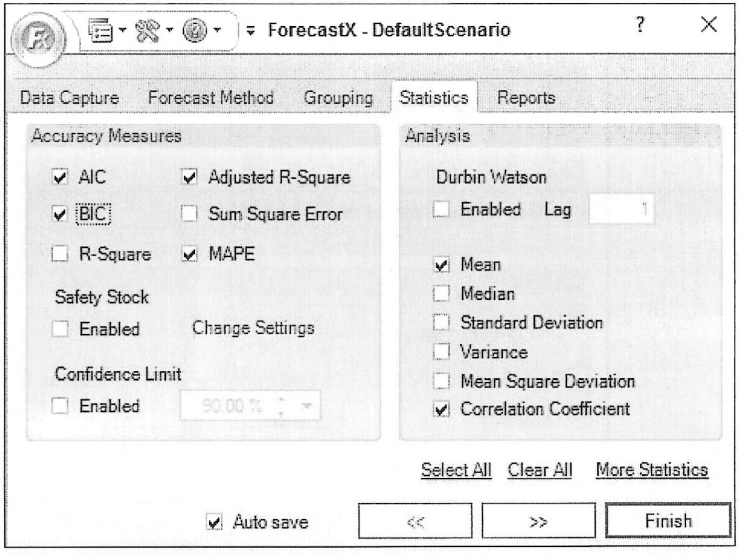
Source: John Galt Solutions

In the **Forecast Method** dialog box, click the down arrow in the **Forecasting Technique** box and select **Multiple Regression.** Make sure the desired variable is selected as the **Dependent Series**, which is **GapSales($M)** in this example. Then click the **Statistics** tab.



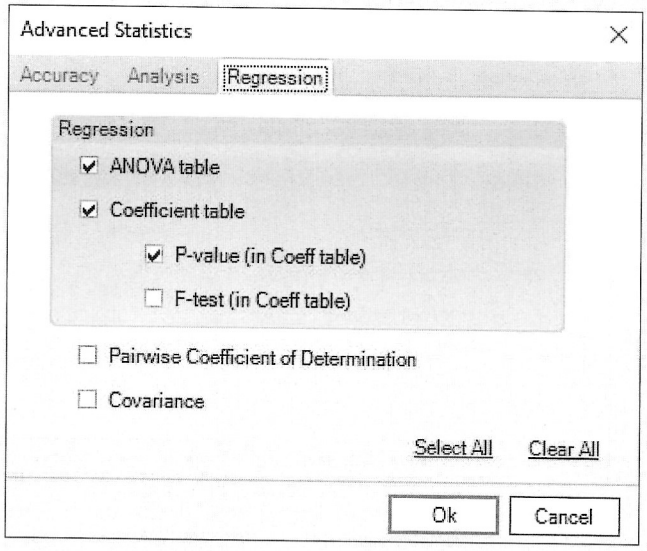
Source: John Galt Solutions

In this dialog box, select the statistics that you desire. Do not forget that there are more choices if you click the **More Statistics** button near the bottom right corner.



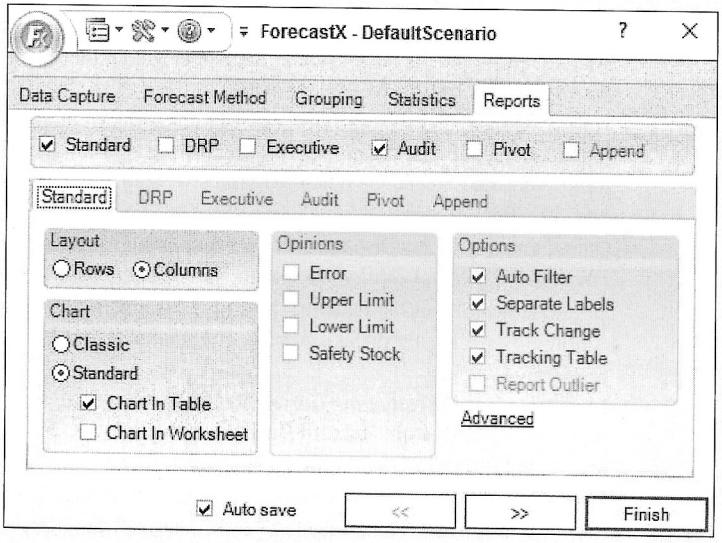
Source: John Galt Solutions

In the More Statistics dialog box, select **Regression** and check the boxes for **ANOVA**, **Coefficients**, and **P-values**. Then click on Ok.

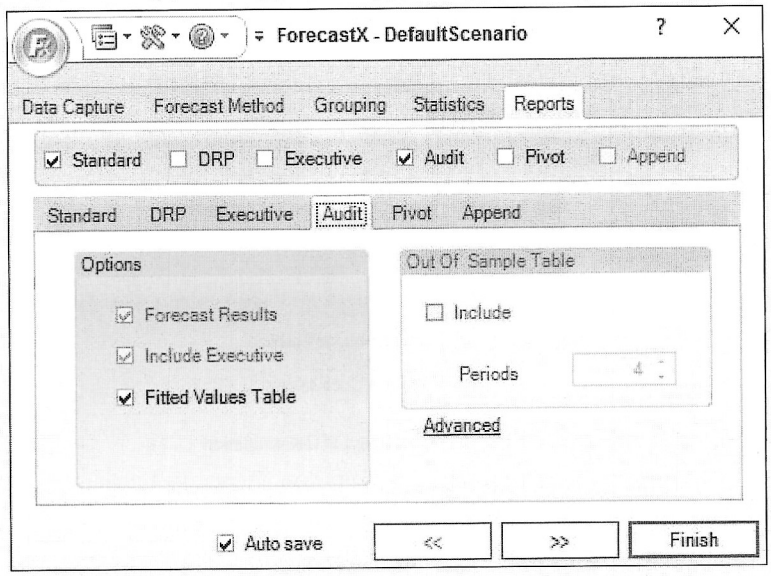


Source: John Galt Solutions

After selecting the statistics you want to see, click the **Reports** tab.
In the **Reports** dialog box, select those you want. Typical selections might be those shown here. If you click the Standard tab, you will want to be sure to select the **Chart in Table** box. In the **Audit Trail** tab, click the **Fitted Values Table**.



Source: John Galt Solutions



Source: John Galt Solutions

Then click the **Finish** button.

ForecastX™ will automatically apply a time-series method to forecast the independent variables. The methods used to forecast the independent variables are shown in the Standard Report. When you have seasonal dummy variables, ForecastX™ will use a Holt-Winters model since the data have a seasonal pattern.

Sometimes you will want to specify the forecast values for some (or all) of the independent variables. This will be illustrated in the appendix to this chapter.

Suggested Readings

Akaike, Hirotugu. "A New Look at Statistical Model Identification." *IEEE Transactions on Automatic Control*, AC-19, 6 (1974).

Aykac, Ahmed; and Antonio Borges. "Econometric Methods for Managerial Applications." In *The Handbook of Forecasting: A Managers Guide.* Eds. Spyros Makridakis and Steven C. Wheelwright. New York: John Wiley & Sons, 1982, pp. 185–203.

Chase, Charles W. Jr. *Demand-Driven Forecasting: A Structured Approach to Forecasting, Second Edition*. Hoboken, New Jersey: John Wiley & Sons, 2013.

Doran, Howard; and Jan Kmenta. "Multiple Minima in the Estimation of Models with Autoregressive Disturbances." *Review of Economics and Statistics* 24 (May 1992), pp. 354–57.

Ellis, Joseph H. *Ahead of the Curve: A Commonsense Guide to Forecasting Business and Market Cycles.* Boston: Harvard Business Review Press, 2005.

Franses, Philip Hans. *Expert Adjustments of Model Forecasts: Theory, Practice and Strategies for Improvement*. Cambridge, U.K.: Cambridge University Press, 2014.

Griffiths, William E.; R. Carter Hill; and George G. Judge. *Learning and Practicing Econometrics*. New York: John Wiley & Sons, 1992.

Gujarati, Damodar N. *Essentials of Econometrics*. New York: McGraw-Hill, 2006.

Jarrell, Stephen B. *Basic Business Statistics.* Boston: Allyn & Bacon, 1988. Especially Chapter 23, "Regression," and Chapter 24, "Evaluating and Forecasting: Two Variable Regression Models."

Johnson, Aaron C., Jr.; Marvin B. Johnson; and Reuben C. Buse. *Econometrics: Basic and Applied.* New York: Macmillan, 1987.

Lewis-Beck, Michael S. *Applied Regression: An Introduction.* Beverly Hills, CA: Sage Publications, 1980.

Mendenhall, William; Robert J. Beaver; and Barbara M. Beaver. *Introduction to Probability and Statistics*. 13th ed. Belmont, CA: Brooks/Cole Cenage Learning, 2009.

Neter, John; William Wasserman; and Michael H. Kutner. *Applied Linear Regression Models.* New York: McGraw-Hill, 1996.

Tetlock, Philip E. and Dan Gardner. *Superforecasting: The Art and Science of Prediction.* New York: Crown Publishers, 2015.

Wallis, K. F. "Testing for Fourth Order Correlation in Quarterly Regression Equations." *Econometrica* 40 (1972), pp. 617–36.

Exercises

1. Explain the difference between bivariate (simple) regression and multiple regression.
2. Explain the five-step process for evaluating a multiple regression model.
3. Describe how a regression plane differs from a regression line.
4. Explain what is meant by a "dummy variable."
5. Describe some ways dummy variables can be used in regression models.
6. Explain things that should be considered when selecting independent variables for a multiple regression model that will be used to make a forecast.
7. Explain why the adjusted R-squared should be used in evaluating multiple-regression models rather than the unadjusted value.
8. The following regression results relate to a study of fuel efficiency of cars as measured by miles per gallon of gas (adjusted R-squared = 0.569; $n = 120$).

Variable*	Coefficient	Standard Error	*t*-Ratio
Intercept	6.51	1.28	
CID	−0.031	0.012	
D	9.46	2.67	
M4	14.64	2.09	
M5	14.86	2.42	
US	−4.64	2.48	

CID = Cubic-inch displacement (engine size)
D = 1 for diesel cars and 0 otherwise
M4 = 1 for cars with a four-speed manual transmission and 0 otherwise
M5 = 1 for cars with a five-speed manual transmission and 0 otherwise
US = 1 for cars made in the United States and 0 otherwise

   a. Calculate the *t*-ratios for each explanatory variable.

   b. Use the first three quick-check regression-evaluation procedures to evaluate this model.

9. Develop a multiple-regression model for auto sales as a function of population and household income from the following data for 10 metropolitan areas:

(c5p9)

Area	Auto Sales (AS) ($000)	Household Income (INC) ($000)	Population (POP) (000)
1	$ 185,792	$ 23,409	133.17
2	85,643	19,215	110.86
3	97,101	20,374	68.04
4	100,249	16,107	99.59
5	527,817	23,432	289.52
6	403,916	19,426	339.98
7	78,283	18,742	89.53
8	188,756	18,553	155.78
9	329,531	21,953	248.95
10	91,944	16,358	102.13

a. Estimate values for b_0, b_1, and b_2 for the following model:

$$AS = b_0 + b_1(INC) + b_2(POP)$$

b. Are the signs you find for the coefficients consistent with your expectations? Explain.

c. Are the coefficients for the two explanatory variables significantly different from zero? Explain.

d. What percentage of the variation in AS is explained by this model?

e. What point estimate of AS would you make for a city where INC = $23,175 and POP = 128.07?

10. In Chapter 4, you worked with data on sales for a line of skiwear that is produced by HeathCo Industries. Barbara Lynch, the product manager for the skiwear, has the responsibility of providing forecasts to top management of sales by quarter one year ahead. One of Ms. Lynch's colleagues, Dick Staples, suggested that unemployment and income in the regions in which the clothes are marketed might be causally connected to sales. If you worked the exercises in Chapter 4, you have developed three bivariate regression models of sales as a function of time (TIME), unemployment (NRUR), and income (INC). Data for these variables and for sales are as follows: (c5p10)

a. Now you can expand your analysis to see whether a multiple-regression model would work well. Estimate the following model:

$$SALES\ b_0 + b_1(INC) + b_2(NRUR)$$

$$SALES = \_\_\_ + / - \_\_\_(INC) + / \_\_\_(NRUR)$$

(Circle + or − as appropriate for each variable)

Do the signs on the coefficients make sense? Explain why.

b. Test to see whether the coefficients you have estimated are statistically different from zero, using a 95 percent confidence level and a one-tailed test.

c. What percentage of the variation in sales is explained by this model?

d. Use this model to make a sales forecast (SF1) for 2017Q1 through 2017Q4, given the previously forecast values for unemployment (NRURF) and income (INCF) as follows:

Period	Sales	Inc	NRUR	Time
Mar-07	72962	218	8.4	1
Jun-07	81921	237	8.2	2
Sep-07	97729	263	8.4	3
Dec-07	142161	293	8.4	4
Mar-08	145592	318	8.1	5
Jun-08	117129	359	7.7	6
Sep-08	114159	404	7.5	7
Dec-08	151402	436	7.2	8
Mar-09	153907	475	6.9	9
Jun-09	100144	534	6.5	10
Sep-09	123242	574	6.5	11
Dec-09	128497	622	6.4	12
Mar-10	176076	667	6.3	13
Jun-10	180440	702	6.2	14
Sep-10	162665	753	6.3	15
Dec-10	220818	796	6.5	16
Mar-11	202415	858	6.8	17
Jun-11	211780	870	7.9	18
Sep-11	163710	934	8.3	19
Dec-11	200135	1010	8	20
Mar-12	174200	1066	8	21
Jun-12	182556	1096	8	22
Sep-12	198990	1162	8	23
Dec-12	243700	1187	8.9	24
Mar-13	253142	1207	9.6	25
Jun-13	218755	1242	10.2	26
Sep-13	225422	1279	10.7	27
Dec-13	253653	1318	11.5	28
Mar-14	257156	1346	11.2	29
Jun-14	202568	1395	11	30
Sep-14	224482	1443	10.1	31
Dec-14	229879	1528	9.2	32
Mar-15	289321	1613	8.5	33
Jun-15	266095	1646	8	34
Sep-15	262938	1694	8	35
Dec-15	322052	1730	7.9	36
Mar-16	313769	1755	7.9	37
Jun-16	315011	1842	7.9	38
Sep-16	264939	1832	7.8	39
Dec-16	301479	1882	7.6	40

Period	**NRURF (%)**	**INC ($ Billions)**	**SF1**
Mar-17	7.6	1,928	_____
Jun-17	7.7	1,972	_____
Sep-17	7.5	2,017	_____
Dec-17	7.4	2,062	_____

e. Actual sales for 2017 were: Q1 = 334,271; Q2 = 328,982; Q3 = 317,921; Q4 = 350,118. On the basis of this information, how well would you say the model worked? What is the mean absolute percentage error (MAPE)?

f. Plot the actual data for 2017Q1 through 2017Q4 along with the values predicted for each quarter based on this model.

11. *a.* Construct a time-series graph of the sales data for HeathCo's line of skiwear (see data in c5p11). Does there appear to be a seasonal pattern in the sales data? Explain why you think the results are as you have found. (c5p11)

b. It seems logical that skiwear would sell better from October through March than from April through September. To test this hypothesis, begin by adding two dummy variables to the data: a dummy variable Q1 = 1 for each first quarter (January, February, March) and Q1 = 0 otherwise; and a dummy variable Q4 = 1 for each fourth quarter (October, November, December) and Q4 = 0 otherwise. Once the dummy variables have been entered into your data set, estimate the following trend model:

$$\text{SALES } b_0 + b_1(\text{TIME}) + b_2\text{Q1} + b_3\text{Q4}$$

Evaluate these results by answering the following:

- Do the signs make sense? Why or why not?
- Are the coefficients statistically different from zero at a 95 percent confidence level (one-tailed test)?
- What percentage of the variation in SALES is explained by this model?

c. Use this model to make a forecast of SALES (SF2) for the four quarters of 2017 and calculate the MAPE for the forecast period.

Period	SALES ($000)	SF2
2017Q1	334,271	_____
2017Q2	328,982	_____
2017Q3	317,921	_____
2017Q4	350,118	_____

d. Prepare a time-series plot of SALES (for 2007Q1 through 2016Q4) along with SF2 (for 2007Q1 through 2017Q4) to illustrate how SALES and SF2 compare.

12. AmeriPlas, Inc., produces 20-ounce plastic drinking cups that are embossed with the names of prominent beers and soft drinks. The sales data are:

Date	Sales	Date	Sales
Jan-13	40,358	Jan-14	37,255
Feb-13	45,002	Feb-14	38,521
Mar-13	63,165	Mar-14	55,110
Apr-13	57,479	Apr-14	51,389
May-13	52,308	May-14	58,068
Jun-13	60,062	Jun-14	64,028
Jul-13	51,694	Jul-14	52,873
Aug-13	54,469	Aug-14	62,584
Sep-13	48,284	Sep-14	53,373
Oct-13	45,239	Oct-14	52,060
Nov-13	40,665	Nov-14	51,727
Dec-13	47,968	Dec-14	51,455

(continued on next page)

(continued)

Date	Sales	Date	Sales
Jan-15	47,906	Jan-16	65,711
Feb-15	53,570	Feb-16	68,005
Mar-15	69,189	Mar-16	78,029
Apr-15	64,346	Apr-16	92,764
May-15	77,267	May-16	97,175
Jun-15	75,787	Jun-16	86,255
Jul-15	74,052	Jul-16	90,496
Aug-15	79,756	Aug-16	87,602
Sep-15	73,292	Sep-16	83,577
Oct-15	77,207	Oct-16	92,610
Nov-15	68,423	Nov-16	73,949
Dec-15	67,274	Dec-16	77,711

(c5p12)

a. Prepare a time-series plot of the sales data. Does there appear to be a regular pattern of movement in the data that may be seasonal? Ronnie Mills, the product manager for this product line, believes that her brief review of sales data for the four-year period indicates that sales are slowest in November, December, January, and February than in other months. Do you agree?

b. Since production is closely related to orders for current shipment, Ronnie would like to have a monthly sales forecast that incorporates monthly fluctuations. She has asked you to develop a trend model that includes a time index and dummy variables for all but the above mentioned four months. Do these results support Ronnie's observations? Explain.

c. Ronnie believes that there has been some increase in the rate of sales growth. To test this and to include such a possibility in the forecasting effort, she has asked that you add the square of the time index (T) to your model (call this new term $T2$). Is there any evidence of increase of sales growth? Compare the results of this model with those found in part (b).

d. Use the model in part (c) to forecast sales for 2017. Calculate the mean absolute percentage error (MAPE) for the first six months of 2017. Actual sales for those six months were:

Jan-2017	87327
Feb-2017	84772
Mar-2017	112499
Apr-2017	102633
May-2017	112996
Jun-2017	119807

Chapter Five Appendix

Combining Forecasts (Ensemble Models)

LEARNING OBJECTIVES

After studying this appendix, you should be able to:

1. Explain why a combination of two forecasts (called ensembles) may be better than either one alone.
2. Explain the process for checking to see if a combination of forecasts would create a bias.
3. Explain how to use regression analysis to select the weights for the forecasts that are being combined.
4. Set up the data table that should be used when combining forecasts.
5. Use ForecastX™ to combine forecasts.

INTRODUCTION

The use of combinations of forecasts has been the subject of a great deal of research in forecasting. An indication of the importance of this concept is the fact that the prestigious *International Journal of Forecasting* had a special section, composed of seven articles, entitled "Combining Forecasts" in the year-end issue of the volume for 1989. In December 1992, an article in the same journal provided strong evidence on the importance of combining forecasts to improve accuracy. It was found that 83 percent of expert forecasters believe that combining forecasts will produce more accurate forecasts than could be obtained from the individual methods!

The idea of combining business forecasting models was originally proposed by Bates and Granger. Since the publication of their article, this strategy has received immense support in almost every empirical test of combined forecasts versus individual uncombined forecasts.

Throughout this book, we have emphasized the use of the mean absolute percentage error (MAPE) as a measure of the effectiveness of ***one*** particular forecasting model. In this appendix, instead of choosing the best model from among