Multiple Regression

Introduction

In this chapter, we continue our study of regression analysis by considering situations involving two or more independent variables.

This subject area, called **multiple regression analysis**, enables us to consider more factors and thus obtain better predictions than are possible with simple linear regression.

Regression Model and Regression Equation

The **multiple regression model** is the equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, \ldots, x_p$ and an error term.

$$\boldsymbol{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon}$$

Where

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters of the model.

ϵ is the error term that accounts for the variability in y that cannot be explained by the linear effect of the p independent variables.

The **multiple regression equation** describes how the mean value of y is related to a set of *p* independent variables, $x_1, x_2, \ldots, x_p$.

$$\boldsymbol{E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}$$

The Estimation Process in Multiple Regression

The *estimated multiple regression equation* is

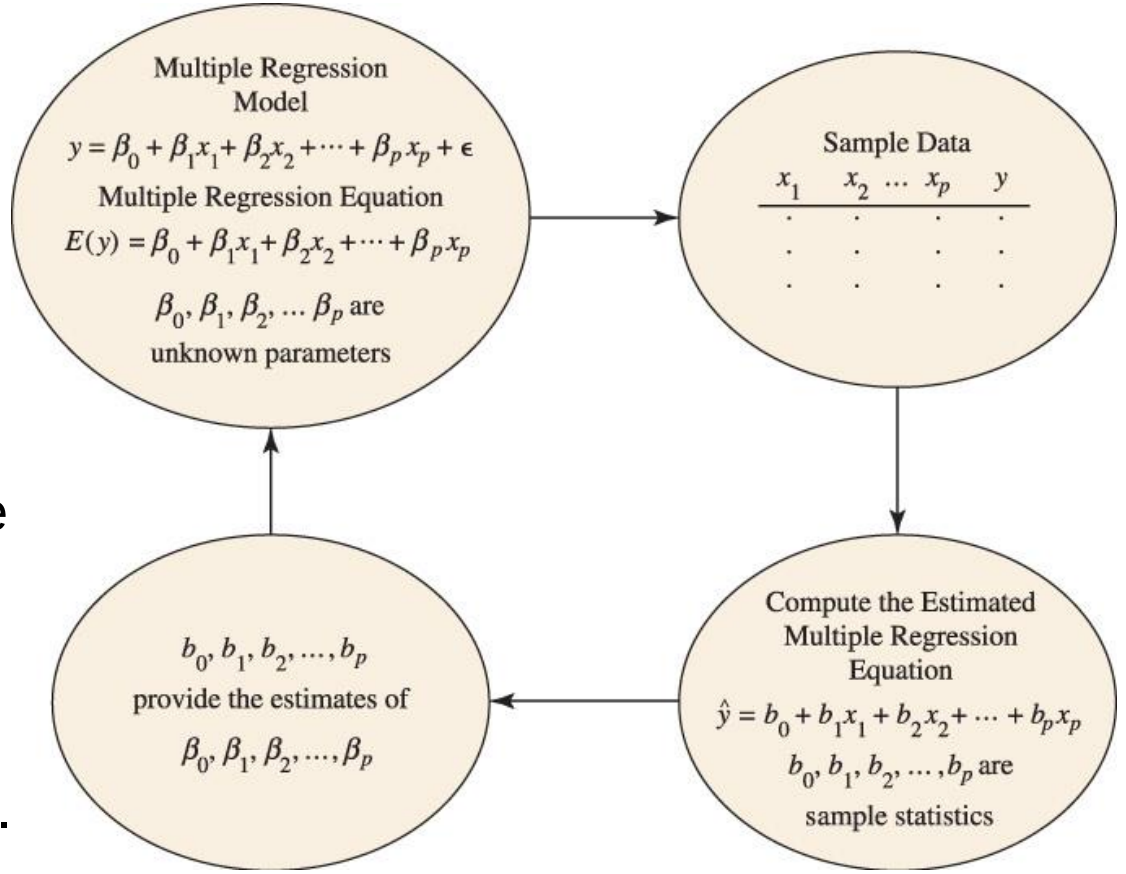
$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

Where:

$\hat{y}$ is a point estimate of $E(y)$ for a given set of *p* independent variables, $x_1, x_2, \ldots, x_p$.

A simple random sample is used to compute the sample statistics $b_0, b_1, b_2, \ldots, b_p$ that are used as estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

The flow chart to the right provides a summary of the estimation process for multiple regression.



Least Squares Method

The **least squares method** uses the sample data to provide the values of the sample statistics $b_0, b_1, b_2, \ldots, b_p$ that minimize the sum of the squares of the deviations between the observed values and the predicted values of the dependent variable.

Least Squares Criterion

$$\mathbf{min} \sum (\boldsymbol{y_i} - \hat{\boldsymbol{y}})^{\mathbf{2}}$$

Where: y_i is the value of dependent variable for the ith observation.

$\hat{y}_i$ is the predicted value of dependent variable for the ith observation.

Because the formulas for the regression coefficients $b_0, b_1, b_2, \ldots, b_p$ involve the use of matrix algebra, we will rely on computer software packages to perform the calculations.

The emphasis will be on how to interpret the computer output rather than on how to make the multiple regression computations.

The Butler Trucking Company Example

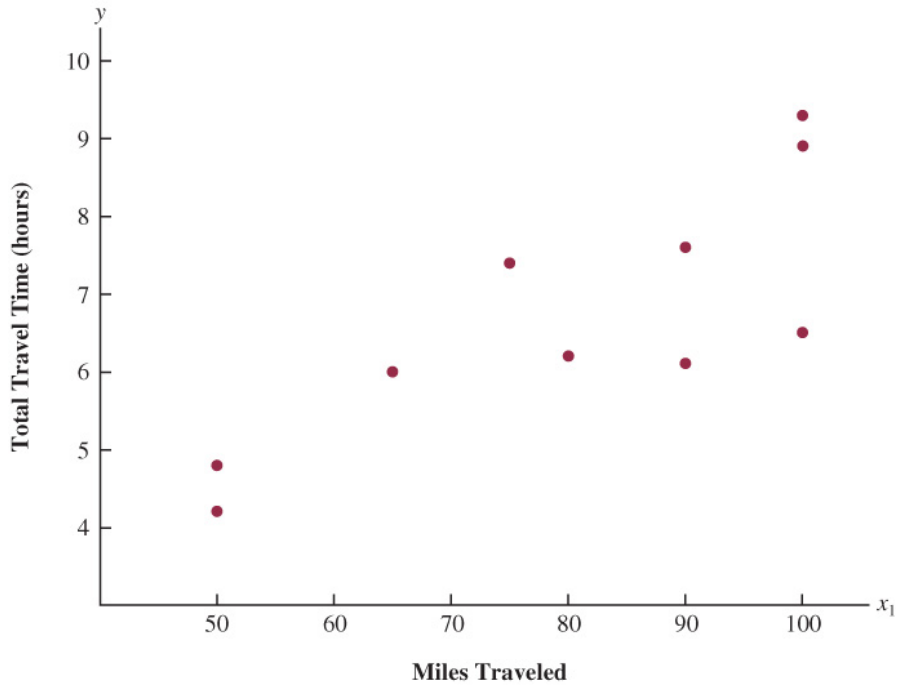
To develop better work schedules, the managers at the Butler Trucking Company, want to predict the total daily travel time for their drivers.

Initially, the managers believed that the total daily travel time (y) would be closely related to the number of miles traveled in making the daily deliveries, x.

DATAfile: *Butler*

The managers used a simple random sample of 10 driving assignments to build a scatter diagram depicting the relationship between the variables.

Because the scatter diagram shows a positive linear relationship, the managers chose the simple linear regression model to represent the relationship between total daily delivery time (y) and the number of miles driven in making daily deliveries (x.)



The Butler Trucking Simple Linear Regression Output with One Independent Variable

The computer output to the right shows the results of the simple linear regression with the total miles traveled as the independent variable x.

$$\hat{y} = 1.27 + 0.0678\,x$$

From the computer output, $F = 15.81$ and $t_1 = 3.98$. Thus, both *F* and *t* tests produce a $p\text{–value} = 0.004$.

At $a = 0.05$, we can reject H_0: $\beta_1 = 0$ and conclude that the relationship between the total travel time and the total miles traveled is significant.

The coefficient of determination also reveals that about 66.41% of the variability in travel time can be explained by the total miles driven.

Regression Statistics	
Multiple R	81.49%
R Square	66.41%
Adj. R Square	62.21%
Standard Error	1.0018
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	1	15.871	15.871	15.81	0.004
Residual	8	8.029	1.004		
Total	9	23.900			

	Coefficients	Std. Err.	*t Stat*	*P-value*
Intercept	1.274	1.401	0.909	0.390
Population	0.06783	0.01706	3.977	0.004

The Butler Trucking Multiple Regression Output with Two Independent Variables

In an attempt to add a second independent variable that explains some of the remaining variability in the dependent variable, the managers felt that the number of deliveries could also contribute to the total travel time.

The computer output to the right shows the results of a multiple regression with both total miles traveled (x_1) and number of deliveries (x_2) as independent variables.

$$\hat{y} = -0.869 + 0.06113x_1 + 0.923x_2$$

Let us now examine the values of the regression coefficients $b_1 = 0.06113$ and $b_2 = 0.923$.

Regression Statistics	
Multiple R	95.07%
R Square	90.38%
Adj. R Square	87.63%
Standard Error	0.5731
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	2	21.601	10.800	32.88	0.000
Residual	7	2.299	0.328		
Total	9	23.900			

	Coefficients	Std. Err.	*t Stat*	*P-value*
Intercept	-0.869	0.952	-0.913	0.392
Miles	0.06113	0.00989	6.182	0.000
Deliveries	0.923	0.221	4.176	0.004

Interpreting Multiple Regression Coefficients

In multiple regression analysis, we interpret each regression coefficient as follows:

> b_i represents an estimate of the change in y corresponding to one unit increase in x_i when all other independent variables are held constant, with $i = 1, \ldots, p$.

In the Butler Trucking example involving two independent variables, we interpret the regression coefficients b_1 and b_2 as follows.

$b_1 = 0.06113$	0.06113 hours is an estimate of the expected increase in travel time corresponding to an increase of one mile in the distance traveled when the number of deliveries is held constant.
$b_2 = 0.923$	0.923 hours is an estimate of the expected increase in travel time corresponding to an increase of one delivery when the number of miles traveled is held constant.

Multiple Coefficient of Determination

The **multiple coefficient of determination**, denoted R^2, indicates that we are measuring the goodness of fit for the estimated multiple regression equation, and it is computed as follows.

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

The partition of SST into SSR + SSE is still the same one we saw in simple linear regression. However, because of its computational difficulty, we rely on computer programs to compute the three sums of squares.

The multiple coefficient of determination can be interpreted as the proportion of the variability in the dependent variable that can be explained by the estimated multiple regression equation.

For the Butler Trucking Company example, we have: $R^2 = \frac{SSR}{SST} = \frac{21.6006}{23.900} = 0.9038$

Thus, about 90.4% of the variability in travel time y is explained by the estimated multiple regression equation with miles traveled and number of deliveries as the independent variables.

Adjusted Multiple Coefficient of Determination

From the previous regression outputs, we noted that the percentage of the variability in travel times that is explained by the estimated regression equation increases from 66.4% to 90.4% when the *number of deliveries* is added as a second independent variable.

In general, R^2 always increases as independent variables are added to the model, because SSE becomes smaller, regardless of whether the added variable is statistically significant.

The **adjusted multiple coefficient of determination**, $\boldsymbol{R_a^2}$, accounts for the effect that the number of independent variables, p, have on the amount of variability explained by the estimated regression equation of a sample with n observations.

$$\boldsymbol{R_a^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-p-1}}$$

For the Butler Trucking Company example, we have:

$$R_a^2 = 1 - (1 - 0.9038)\frac{10-1}{10-2-1} = 0.8763$$

Graphical Representation of a Multiple Regression Equation with Two Independent Variables

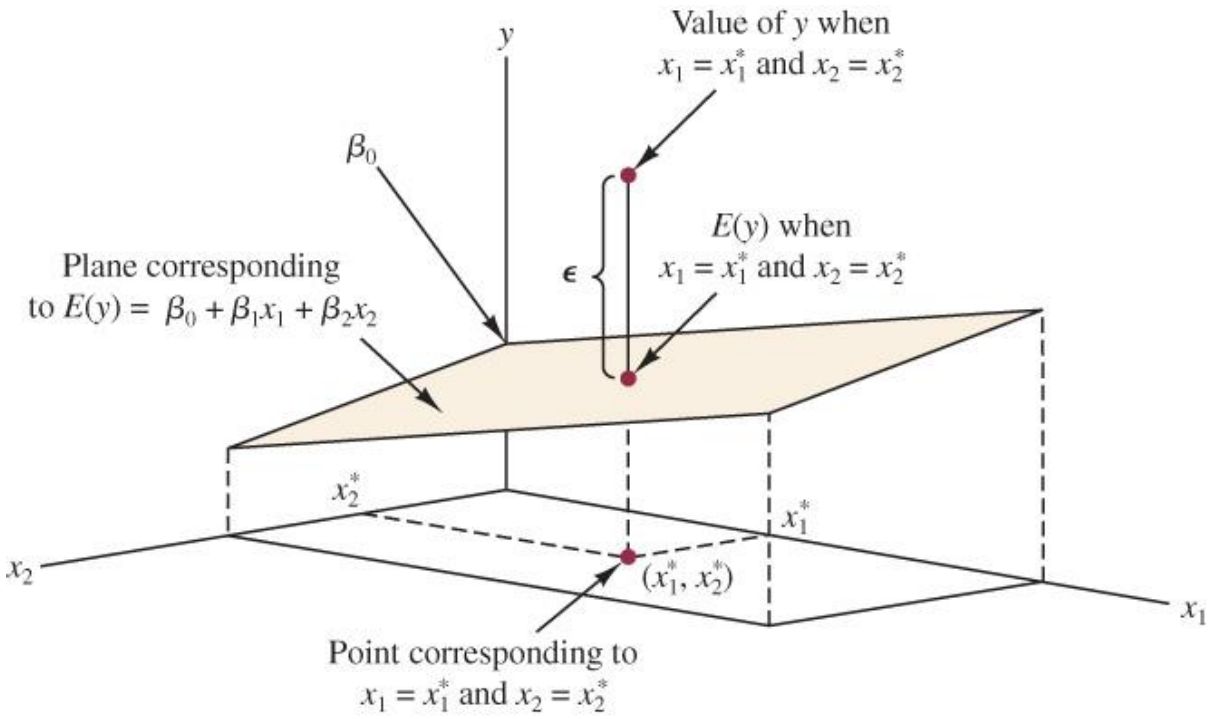
To obtain more insight about the form of the relationship given by a multiple regression equation, consider the following two-independent-variable version.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

The term *response variable* is often used in place of the term *dependent variable*.

The graph of this equation represented in three-dimensional space is a plane called a *response surface*.

Note that the value of ϵ shown in the graph is the difference between the actual y value and the expected value of y, $E(y)$, when $x_1 = x_1^*$ and $x_2 = x_2^*$.



Assumptions About the Error Term ϵ

In the multiple regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$, we have the following assumptions about the error term ϵ:

1. The error term ϵ is a random variable with an expected value of zero; that is, $E(\epsilon) = 0$.

   Implication: for given values of $x_1, x_2, \ldots, x_p$, the multiple regression equation is given by

   $$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p.$$

2. The variance of ϵ, σ^2, is the same for all values of the independent variables $x_1, x_2, \ldots, x_p$.

   Implication: The variance of y about the regression line equals σ^2 for all the values of $x_1, x_2, \ldots, x_p$.

3. The values of ϵ are independent.

   Implication: The value of ϵ for a particular set of values for the independent variables is not related to the value of ϵ for any other set of values.

4. The error term ϵ is a normally distributed random variable reflecting the deviation between y and $E(y)$.

   Implication: Because $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are constant for the given values of $x_1, x_2, \ldots, x_p$, the dependent variable y is also a normally distributed random variable.

Testing for Significance in Multiple Regression

In simple linear regression, the F and t test provide the same conclusion.

However, in multiple regression, the F and t test have different purposes.

1. The F test is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables.

   We will refer to the F test as the test for *overall significance*.

2. If the F test shows an overall significance, the t test is used to determine whether each of the individual independent variables is significant.

   A separate t test is conducted for each of the independent variables in the model.

   We refer to each of these t tests as a test for *individual significance*.

In the material that follows, we will explain the F and the t test and apply each to the Butler Trucking Company example.

F Test for Overall Significance

With *p* independent variables, the hypotheses for the F test are set up as follows.

$$\boldsymbol{H_0}: \boldsymbol{\beta_1 = \beta_2 = \ldots = \beta_p = 0}$$

$$\boldsymbol{H_a}: \textbf{At least one parameter is not equal to zero}$$

The test statistic follows an F distribution with *p* degree of freedom at the numerator, and $n - p - 1$ degrees of freedom at the denominator.

$$\boldsymbol{F = \frac{MSR}{MSE}}$$

Where, $MSR = SSR/p$

$$MSE = SSE/(n - p - 1)$$

The rejection rule is as follows

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

Critical value approach: Reject H_0 if $F \geq F_\alpha$

F Test for the Butler Trucking Example

The hypotheses for the F test in the Butler trucking multiple regression example with miles traveled (x_1) and number of deliveries (x_2) as the two independent variables are.

H_0: $\beta_1 = \beta_2 = 0$

H_a: β_1 and/or β_2 is not equal to zero

In the analysis of variance computer output from the previous section, we found

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{21.6006/2}{2.2994/(10-2-1)} = \frac{10.8003}{0.3285} = 32.88$$

The rejection rule, using a level of significance $\alpha = 0.01$, is as follows.

p-value approach: $p\text{–value} = 0.000 \leq 0.01 = \alpha$. Reject H_0

Critical value approach: $F = 32.88 \geq F_{.01} = 9.55$. Reject H_0

We conclude that a significant relationship is present between travel time (y) and the two independent variables, miles traveled (x_1) and number of deliveries (x_2) .

Estimate of σ^2 in Multiple Regression

From the assumptions on the error term, we concluded that σ^2, the variance of ϵ, also represents the variance of the residuals; that is, the deviations deviation between y and $E(y)$.

In multiple regression, the **mean square error** (MSE) also provides an unbiased estimate of σ^2.

$$\boldsymbol{s^2 = MSE = SSE/(n - p - 1)}$$

To estimate the standard deviation of ϵ, we take the square root of s^2. The resulting value, *s*, is referred to as the **standard error of the estimate**.

$$\boldsymbol{s = \sqrt{MSE} = \sqrt{SSE/(n - p - 1)}}$$

From the Butler Trucking multiple regression example computer output, we found $\mathrm{M}SE = 0.3285$, Thus, we have

$$s = \sqrt{MSE} = \sqrt{0.3285} = 0.5731$$

Note that the value of the standard error of the estimate also appears in the computer output.

ANOVA Table for Multiple Regression

The ANOVA table for multiple regression displays the results of the F test.

Source of Variation	Sum of Squares	Deg. of Freedom	Mean Squares	*F*	*p*-value
Regression	SSR	p	MSR = SSR / p	MSR / MSE	$Pr(F \geq \text{MSR} / \text{MSE})$
Error	SSE	$n - p - 1$	MSE = SSE / $(n - p - 1)$		
Total	SST	$n - 1$			

Regression, Error, and Total are the labels for the three sources of variation, with SSR, SSE, and SST appearing as the corresponding sum of squares in the second column.

The ANOVA table with the F test computations for the Butler Trucking example follows.

Source of Variation	Sum of Squares	Deg. of Freedom	Mean Squares	*F*	*p*-value
Regression	21.6006	2	10.8003	32.88	0.0003
Error	2.2994	7	0.3285		
Total	23,9000	9			

t Test for Individual Significance

In a multiple regression model with *p* independent variables, for any parameter β_i, and $i = 1, \ldots, p$, we have the following hypotheses.

$$\boldsymbol{H_0: \beta_i = 0}$$

$$\boldsymbol{H_a: \beta_i \neq 0}$$

The test statistic follows a *t* distribution with $n - p - 1$ degrees of freedom.

$$\boldsymbol{t = \frac{b_i}{s_{b_i}}}$$

Where s_{b_i} is the estimated standard deviation of the sample regression coefficient b_i.

The rejection rule is as follows.

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

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

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

Individual *t* Tests for the Butler Trucking Example

The multiple regression output for the Butler Trucking problem shows the *t*-ratio calculations. The values of b_1, b_2, s_{b_1}, and s_{b_1} are as follows.

$$b_1 = 0.06113 \qquad s_{b_1} = 0.00989$$

$$b_2 = 0.923 \qquad s_{b_1} = 0.221$$

Calculation of the *t*-ratios provide the test statistic for the hypotheses involving parameters β_1 and β_2, also provided by the computer output.

$$b_1/s_{b_1} = 0.06113/0.00989 = 6.18$$

$$b_2/s_{b_2} = 0.923/0.221 = 4.18$$

Using $\alpha = 0.01$, the *p*-values of 0.000 and 0.004 in the output indicate that we can reject H_0: $b_1 = 0$ and H_0: $b_2 = 0$. Hence, both parameters are statistically significant.

Alternatively, Table 2 of Appendix B shows that with $n - p - 1 = 7$ degrees of freedom, $t_{.005} = 3.499$. Because $6.18 \geq 3.499$ and $4.18 \geq 3.499$, we reject H_0: $b_1 = 0$ and H_0: $b_2 = 0$.

Multicollinearity

Consider a modification of the Butler Trucking example in which the independent variables are total miles driven (x_1) and gas consumption (x_2.)

The F test shows the overall significance of the model, but the individual *t* tests fail to reject H_0.

This is because gas consumption depends on the number of miles traveled. In fact, the two variables have a strong positive correlation ($r = 0.9586$.)

Multicollinearity refers to the correlation among the independent variables.

If the estimated regression equation is used only for predictive purposes, multicollinearity is usually not a serious problem, but every attempt should be made to avoid including independent variables that are highly correlated ($|r| \geq 0.7$.)

Regression Statistics	
Multiple R	84.03%
R Square	70.61%
Adj. R Square	62.21%
Standard Error	1.0017
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	2	16.876	8.438	8.41	0.014
Residual	7	7.024	1.003		
Total	9	23.900			

	Coefficients	*Std. Err.*	*t Stat*	*P-value*
Intercept	1.681	1.458	1.152	0.287
Miles	0.1253	0.0599	2.091	0.075
Consumption	-0.615	0.615	-1.001	0.350

Using the Estimated Regression Equation for Estimation and Prediction

The procedures for estimating the mean value of y and predicting an individual value of y in multiple regression are similar to those we saw in simple linear regression.

We substitute the given values of $x_1, x_2, \ldots, x_p$ into the estimated regression equation and use the corresponding value of $\hat{y}$ as the point estimate.

If we use the estimated regression equation for a truck that traveled 100 miles ($x_1 = 100$) to make two deliveries ($x_2 = 2$), we obtain the following point estimate of the total travel time.

$$\hat{y} = -0.869 + 0.06113x_1 + 0.923x_2 = -0.869 + 0.06113(100) + 0.923(2) = 7.09$$

The formulas required to develop interval estimates for the mean value of y and for an individual value of y are beyond the scope of the textbook, but statistical software for multiple regression will often provide these interval estimates.

In the next slide, we show the 95% confidence and prediction intervals for each of the ten observations included in the Butler Trucking data set.

The 95% Confidence and Prediction Intervals for the Butler Trucking Example

Note that the interval estimate for an individual value of y is wider than the interval estimate for the expected value of y because we can always estimate the mean travel time for all trucks with more precision than we can predict the travel time for one specific truck.

Value of x_1	Value of x_2	95% Confidence Interval		95% Prediction Interval	
		Lower Limit	Upper Limit	Lower Limit	Upper Limit
100	4	8.135	9.742	7.363	10.514
50	3	4.127	5.789	3.369	6.548
100	4	8.135	9.742	7.363	10.514
100	2	6.258	7.925	5.500	8.683
50	2	3.146	4.924	2.414	5.656
80	2	5.232	6.505	4.372	7.366
75	3	6.037	6.936	5.059	7.915
65	4	5.960	7.637	5.205	8.392
90	3	6.917	7.891	5.964	8.844
90	2	5.776	7.184	4.953	8.007
75	4	6.669	8.152	5.865	8.955

Dummy Variables

Thus far, the regression examples we have considered involved quantitative independent variables such as student population, distance traveled, and number of deliveries.

However, in many situations, we must work with **categorical independent variables**, such as:

gender (male, female)

method of payment (cash, credit card, check)

To incorporate a categorical independent variable with two levels, such as gender, into a regression model, we define a **dummy variable**, also called an *indicator variable*, as follows.

$$x_i = \begin{cases} 0 \text{ if the gender is female} \\ 1 \text{ if the gender is male} \end{cases}$$

More complex categorical variable with *k* levels, where $k > 2$, such as in the case of the categorical variable method of payment, must be modeled using $k - 1$ dummy variables.

An Example: Johnson Filtration, Inc.

Managers at Johnson Filtration, Inc. want to predict the repair time necessary for each maintenance request using information from past service calls that include the number of months since the last service and the type of repair, whether electrical or mechanical.

DATAfile: *Johnson*

we develop a simple linear regression model using data from a sample of 10 service calls:

dependent variable y: repair time in hours

independent variable x: # of months since last service request.

At $\alpha = 0.05$ significance level, *t* (or F) test indicates that x and y are significantly related.

The coefficient of determination indicates that the number of months since the last service alone explains 53.42% of the variability in repair time.

Regression Statistics	
Multiple R	73.09%
R Square	53.42%
Adj. R Square	47.59%
Standard Error	0.7810
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	1	5.596	5.596	9.174	0.016
Residual	8	4.880	0.610		
Total	9	10.476			

	Coeffs.	*Std. Err.*	*t Stat*	*P-value*
Intercept	2.147	0.605	3.549	0.008
# Months	0.304	0.100	3.029	0.016

Use of a Dummy Variable in the Johnson Filtration Example

To incorporate the type of repair into the regression model, we define the dummy variable:

$$x_2 = \begin{cases} 0 \text{ if the type of repair is mechanical} \\ 1 \text{ if the type of repair is electrical} \end{cases}$$

Note that in this multiple regression model, the independent variable number of months since last service has been redubbed x_1.

At $\alpha = 0.05$ significance level, the F test indicates that the model is overall significant, and that both x_1 and x_2 are individually significant.

The coefficient of determination indicates that the two independent variables explain 85.92% of the variability in repair time.

Regression Statistics	
Multiple R	92.69%
R Square	85.92%
Adj. R Square	81.90%
Standard Error	0.4590
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*P-value*
Regression	2	9.001	4.500	21.36	0.0010
Residual	7	1.475	0.211		
Total	9	10.476			

	Coeffs	*Std Err.*	*t Stat*	*P-value*
Intercept	0.930	0.467	1.993	0.087
# Months	0.388	0.063	6.195	0.000
Repair Type	1.263	0.314	4.020	0.005

Interpreting the Parameter of a Dummy Variable

The multiple regression equation for the Johnson Filtration example is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

To interpret the parameters β_0, β_1, and β_2, when a categorical variable is present, consider the case when $x_2 = 0$ (mechanical repair.)

We use $E(y \mid \text{mechanical})$ to denote the mean of repair time given a mechanical repair ($x_2 = 0$), and $E(y \mid \text{electrical})$ for the corresponding mean given an electrical repair ($x_2 = 1$).

$$E(y \mid \text{mechanical}) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1$$

$$E(y \mid \text{electrical}) = \beta_0 + \beta_1 x_1 + \beta_2(1) = \beta_0 + \beta_1 x_1 + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x_1$$

The slope of both equations is β_1, but the y-intercept is β_0 for $E(y \mid \text{mechanical})$ and $\beta_0 + \beta_2$ for $E(y \mid \text{electrical})$.

β_2 indicates the difference in mean repair time between a mechanical and an electrical service, the two levels of the independent categorical variable type of repair.

Interpreting the Parameters in the Johnson Filtration Example

If β_2 is positive, the mean repair time for an electrical repair will be greater than that for a mechanical repair; the opposite will be true if β_2 is negative. If $\beta_2 = 0$, there is no difference.

Given, $b_0 = 0.930$ and $b_2 = 1.263$, we have

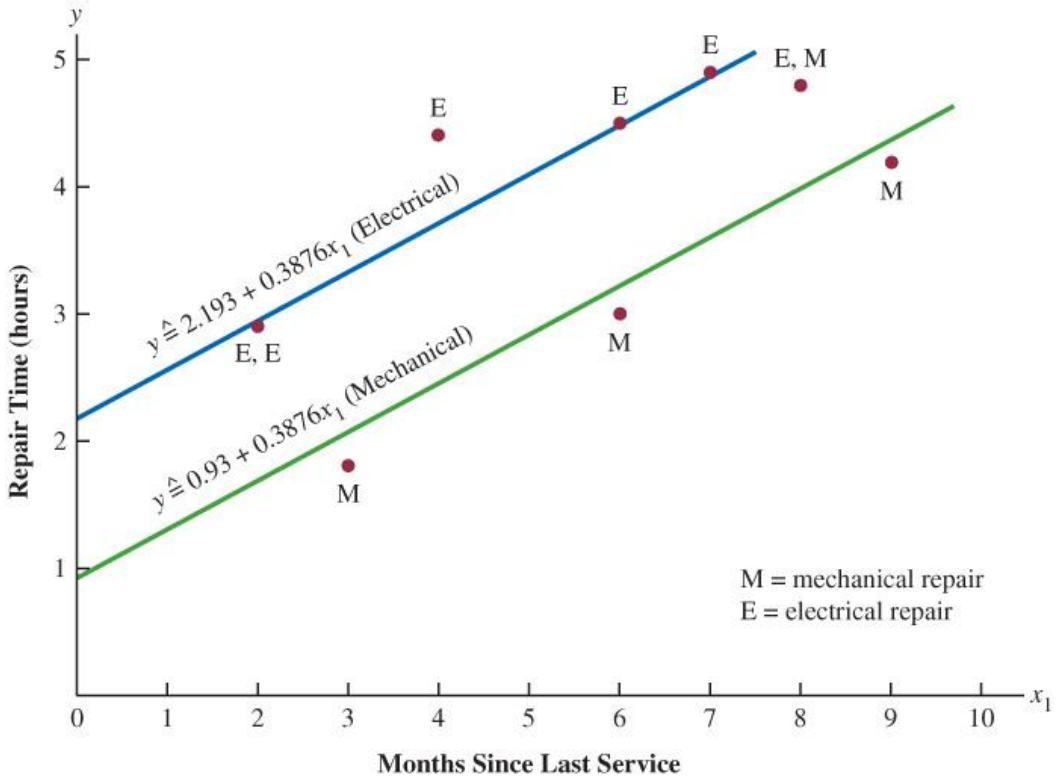
For mechanical repairs, $x_2 = 0$, and

$$\hat{y} = 0.930 + 0.3876x_1$$

For electrical repairs, $x_2 = 1$, and

$$\hat{y} = (0.930 + 1.263) + 0.3876x_1$$
$$= 2.193 + 0.3876x_1$$

The two equations form parallel lines with a y-intercept difference equal to $b_2 = 1.263$. Thus, on average, electrical repairs require 1.263 hours longer than mechanical repairs.



More Complex Categorical Variables

If an independent categorical variable has k levels, $k-1$ dummy variables are required, with each dummy variable being coded as 0 or 1.

Consider the situation faced by a manufacturer of copy machines that sells its products to three sales territories: region A, B, and C.

To code the sales regions in a multiple regression model that explains the dependent variable (y) number of units sold, we need to define $k-1=2$ dummy variables as follows.

$$x_1 = \begin{cases} 1 \text{ if sales region B} \\ 0 \text{ otherwise} \end{cases}$$

$$x_2 = \begin{cases} 1 \text{ if sales region C} \\ 0 \text{ otherwise} \end{cases}$$

Sales Region	x_1	x_2
A	0	0
B	1	0
C	0	1

The regression equation relating the expected number of units sold to the sales region can be written as $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

Interpretation of the Parameters for a Categorical Variable with Three Levels

To interpret β_0, β_1, and β_2, in the sales territory example, consider the following variations of the regression equation.

$$E(y \mid \text{region A}) = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$$

$$E(y \mid \text{region B}) = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$$

$$E(y \mid \text{region C}) = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$$

Thus, the regression parameters are interpreted as follows.

β_0 is the mean or expected value of sales for region A.

β_1 is the difference between the mean number of units sold in region B and the mean number of units sold in region A.

β_2 is the difference between the mean number of units sold in region C and the mean number of units sold in region A.

Standardized Residuals in Multiple Regression

In Chapter 14, we introduced the formula for standardized residual for observation i as

$$\frac{y_i - \hat{y}_i}{s_{y_i - \hat{y}_i}}$$ where $s_{y_i - \hat{y}_i} = s\sqrt{1 - h_i}$ is the standard deviation for residual i

h_i is the **leverage** for observation i and is determined by how far the values of the independent variables are from their means.

Because the computation of leverage values in multiple regression analysis is too complex to be done by hand, we use statistical software to compute the standardized residuals.

The standardized residuals for the Butler Trucking example are shown.

Miles Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Predicted Value ($\hat{y}$)	Residual ($y - \hat{y}$)	Standardized Residual
100	4	9.3	8.93846	0.361541	0.78344
50	3	4.8	4.95830	-0.158304	-0.34962
100	4	8.9	8.93846	-0.038460	-0.08334
100	2	6.5	7.09161	-0.591609	-1.30929
50	2	4.2	4.03488	0.165121	0.38167
80	2	6.2	5.86892	0.331083	0.65431
75	3	7.4	6.48667	0.913331	1.68917
65	4	6.0	6.79875	-0.798749	-1.77372
90	3	7.6	7.40369	0.196311	0.36703
90	2	6.1	6.48026	-0.380263	-0.77639

Standardized Residual Plot for Butler Trucking

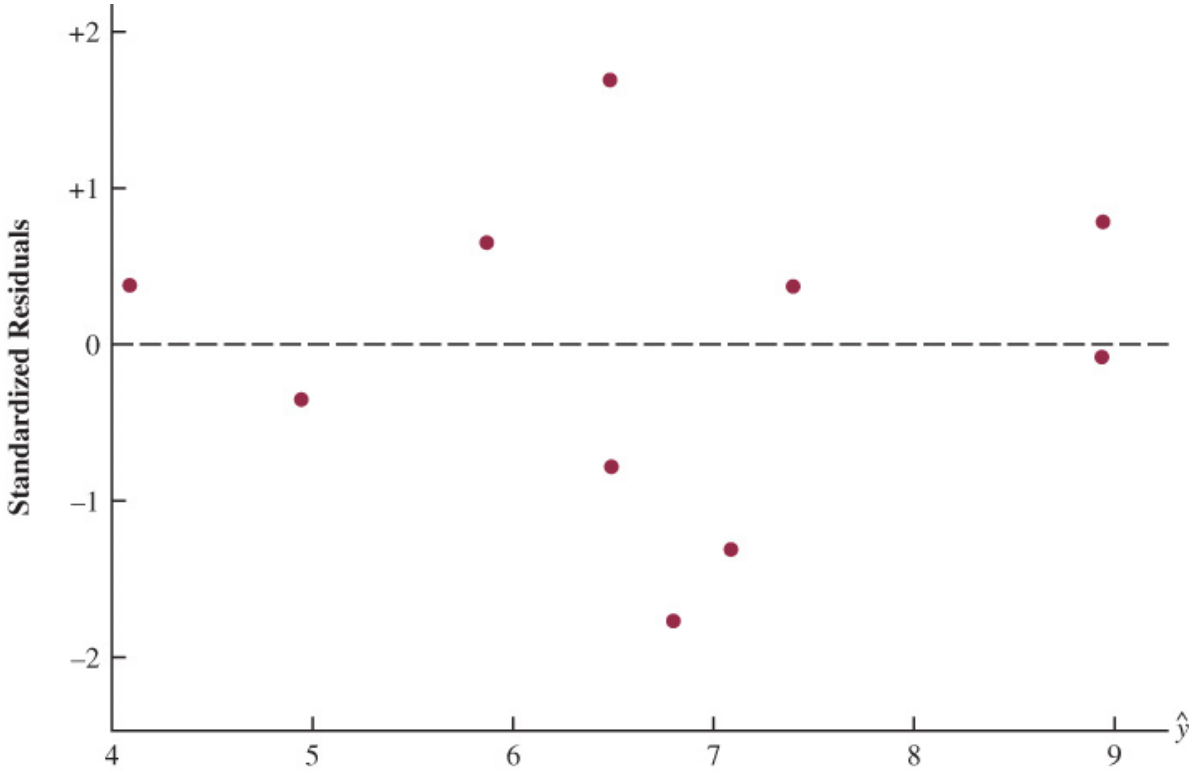
The standardized residual plot against $\hat{y}$ for the Butler Trucking multiple regression example shown below does not indicate any unusual abnormalities.

Because all the standardized residuals are between −2 and +2, we have no reason to question the assumption that the error term ϵ is normally distributed,

Thus, we conclude that the model assumptions are reasonable.

A normal probability plot also can be used to determine whether the distribution of ϵ appears to be normal.

The same procedure to build a normal probability plot shown in Chapter 14 is also appropriate for multiple regression.



Detecting Outliers in Multiple Regression

An **outlier** is a data point (observation) that does not fit the trend shown by the remaining data.

> *An observation is classified as an outlier if the value of its standardized residual is less than −2 or greater than +2.*

Applying this rule to the data set for the Butler Trucking example, we do not detect any outliers.

The presence of one or more outliers in a data set tends to increase s, the standard error of the estimate, and consequently increase $s_{y_i-\hat{y}_i}$, the standard deviation of residual i.

Because $s_{y_i-\hat{y}_i}$ appears in the denominator of the formula for the standardized residual, the size of the standardized residual will decrease as s increases, making it harder to identify the observation as being an outlier.

We can circumvent this difficulty by using **studentized deleted residuals**, defined as follows:

> *The studentized deleted residual for observation i is a standardized residual computed using the standard deviation of residual i for a data set of $n - 1$ observations in which observation i was deleted from the original data set of n observations.*

Studentized Deleted Residuals and Outliers

If observation i is an outlier, $s_{(i)}$, the standard error of the estimate computed with the deleted observation i, will be less than s, the standard error of the estimate for the original set.

Thus, if observation i is an outlier, the absolute value of the ith studentized deleted residual will be larger than the corresponding standardized residual.

The studentized deleted residuals for the Butler example are shown to the right.

Because for the calculations we used $n - 1 = 10 - 1 = 9$ observations, then SSE follows a t distribution with $(n - 1) - p - 1 = 9 - 2 - 1 = 6$ degrees of freedom.

Because observation i is an outlier if the absolute value of the studentized deleted residual is greater than $t_{.025} = 2.447$, it follows that we do not detect any outliers.

Miles Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Standardized Residual	Studentized Deleted Residual
100	4	9.3	0.78344	0.75939
50	3	4.8	-0.34962	-0.32654
100	4	8.9	-0.08334	-0.07720
100	2	6.5	-1.30929	-1.39494
50	2	4.2	0.38167	0.35709
80	2	6.2	0.65431	0.62519
75	3	7.4	1.68917	2.03187
65	4	6.0	-1.77372	-2.21314
90	3	7.6	0.36703	0.34312
90	2	6.1	-0.77639	-0.75190

Influential Observations in Multiple Regression

Leverage of observation i, denoted h_i, measures how far the values of the independent variables are from their mean values.

Thus, leverage is used to identify observations for which the value of the independent variable have a strong influence on the regression results.

We use the rule of thumb $h_i > 3\,(p+1)/n$ to identify influential observations.

For the Butler Trucking example with $p = 2$ independent variables and $n = 10$ observations, the critical value for leverage is

$$3\,(p+1)/n = 3\,(2+1)/10 = 0.9$$

As shown to the right, because h_i does not exceed 0.9 for any of the observations, we do not detect influential observations in the data set.

Miles Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Leverage
100	4	9.3	0.351704
50	3	4.8	0.375863
100	4	8.9	0.351704
100	2	6.5	0.373451
50	2	4.2	0.430220
80	2	6.2	0.220557
75	3	7.4	0.110009
65	4	6.0	0.382657
90	3	7.6	0.129098
90	2	6.1	0.269737

Cook's Distance for Influential Observations

Because an observation can be identified as having high leverage and not necessarily be influential in terms of the resulting estimated regression equation, we can use Cook's distance to identify influential observations in terms of both leverage and residual.

The **Cook's distance measure** for observation i is defined as

$$D_i = \frac{(y_i - \hat{y}_i)^2}{(p+1)s^2}\left[\frac{h_i}{(1-h_i)^2}\right]$$

We use the rule of thumb $D_i > 1$ to identify whether observation i is influential.

The calculations of Cook's distance for the Butler Trucking example are shown to the right.

Because there are no observations with $D_i > 1$, we do not detect any influential observation.

Miles Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Leverage	Cook's D
100	4	9.3	0.351704	0.110994
50	3	4.8	0.375863	0.024536
100	4	8.9	0.351704	0.001256
100	2	6.5	0.373451	0.347923
50	2	4.2	0.430220	0.036663
80	2	6.2	0.220557	0.040381
75	3	7.4	0.110009	0.117562
65	4	6.0	0.382657	0.650029
90	3	7.6	0.129098	0.006656
90	2	6.1	0.269737	0.074217

Logistic Regression Equation

Logistic regression expands the application of regression analysis to applications in which the dependent variable may only assume two discrete values, such as 0 and 1.

The **logistic regression equation** is a nonlinear equation describing how the mean value of a discrete variable y is related to a set of *p* independent variables, $x_1, x_2, \ldots, x_p$.

$$\boldsymbol{E(y) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p}}}$$

Where

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are the parameters of the model.

y is a discrete dependent variable whose values are coded 0 and 1.

Because the value of $E(y)$ is interpreted as the probability that $y = 1$ given a particular set of values for the independent variables $x_1, x_2, \ldots, x_p$,we write the logistic regression equation as

$$\boldsymbol{E(y) = P(y = 1|x_1, x_2, \ldots, x_p)}$$

Graphical Interpretation of Logistic Regression

To better visualize the logistic regression equation, consider a simple model involving only one independent variable x, with the values of the model parameters being $\beta_0 = -7$ and $\beta_1 = 3$.

The logistic regression equation corresponding to these parameter values is

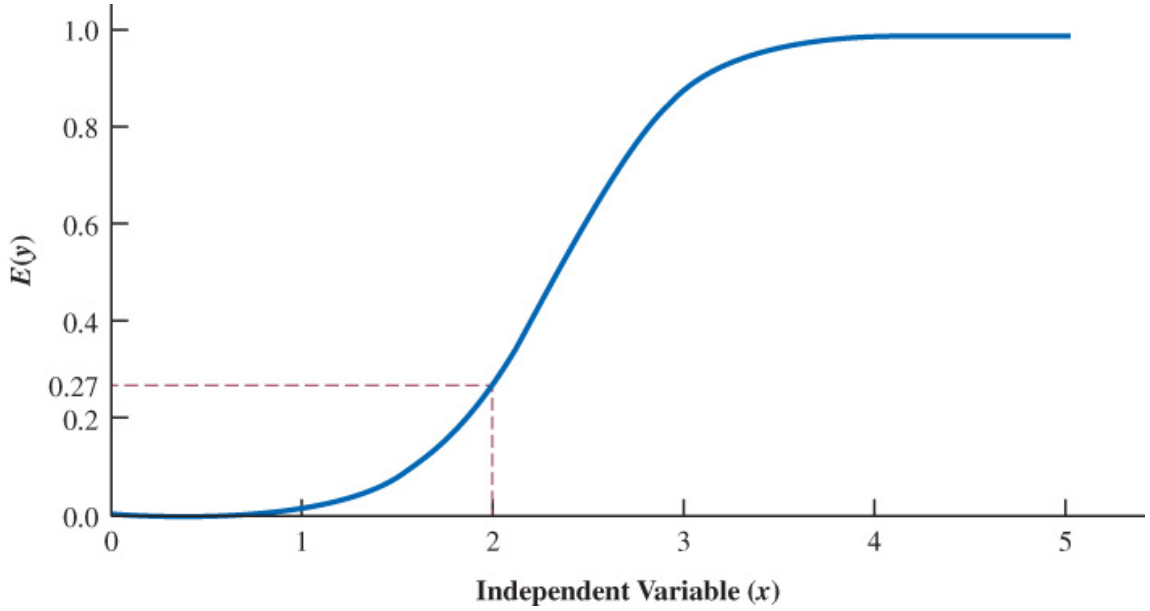
$$E(y) = P(y = 1|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} = \frac{e^{-7+3x}}{1 + e^{-7+3x}}$$

Note that the graph is S-shaped, with the value of $E(y)$ ranging from 0 to 1.

For example, when $x = 2$, $E(y)$ is about 0.27.

Note also that $E(y)$

- approaches 1 for larger values of x
- approaches 0 as x becomes smaller
- increases rapidly as x goes from 2 to 3.



Estimated Logistic Regression Equation and Testing for Significance

The **estimated logistic regression equation** used to provide an estimate of the probability that $y = 1$ given a set of p independent variables, $x_1, x_2, \ldots, x_p$, is written as follows.

$$\hat{y} = \text{estimate of } P(y = 1 | x_1, x_2, \ldots, x_p) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p}}$$

Where $b_0, b_1, b_2, \ldots, b_p$ are the estimated logistic regression coefficients.

Testing for significance in logistic regression is similar to testing for significance in multiple regression. The hypotheses are set the same way as we did for multiple regression analysis.

The only difference is that in the test for overall significance, the X^2 test statistic is a random variable following a chi-square distribution with p degrees of freedom.

The X^2 test statistic for the individual significance of the independent variables also follows a chi-square distribution with 1 degree of freedom.

An Application of Logistic Regression

Simmons Stores, owner and operator of a national chain of women's apparel stores, would like to send a limited edition of an expensive sales catalog to only those customers who have a high probability of using the included coupon.

Management wants to predict whether a customer receiving the catalog will redeem the coupon (dependent variable y) using a logistic regression model that has two independent variables:

annual spending at Simmons stores (x_1)

whether a customer has a Simmons credit card (x_2)

Simmons selected a simple random sample of 50 Simmons card customers and another simple random sample of 50 customers who do not have a Simmons card (DATAfile: *Simmons.)*

Simmons sent the catalog to each of the 100 selected customers, and at the end of a test period, noted whether each customer had used their coupon (coding 1 if used and 0 if not), and recorded the amount spent in thousands of dollars by each customer over the previous year.

Logistic Regression for the Simmons Stores Example

The variables are defined as:

$$y = \begin{cases} 0 \text{ if customer did not use coupon} \\ 1 \text{ if customer used coupon} \end{cases}$$

$$x_1 = \text{annual spending at Simmons stores (\$,1000)}$$

$$x_2 = \begin{cases} 0 \text{ if customer does not have Simmons card} \\ 1 \text{ if customer has Simmons card} \end{cases}$$

Using the estimates for β_0, β_1, and β_2 from the computer output, the estimated logistic regression equation is

$$\hat{y} = \frac{e^{b_0+b_1x_1+b_2x_2}}{1+e^{b_0+b_1x_1+b_2x_2}} = \frac{e^{-2.146+0.342x_1+1.099x_2}}{1+e^{-2.146+0.342x_1+1.099x_2}}$$

We can now use the equation to estimate the probability of using the coupon for a particular type of customer.

Significance Tests

Term	*df*	X^2	*P-value*
Whole Model	2	13.63	0.0011
Spending	1	7.56	0.0060
Card	1	6.41	0.0013

Parameter Estimates

Term	*Estimate*	*Standard Error*
Intercept	-2.146	0.577
Spending	0.342	0.13
Card	1.099	0.440

Odds Ratios

Term	*Odds Ratio*	*Lower 95%*	*Upper 95%*
Spending	1.4073	1.0936	1.8109
Card	3.0000	1.2550	7.1730

Estimating Probability with Logistic Regression

Suppose we have two groups of Simmons customers who spend $2,000 annually, but only one of the groups uses the Simmons card for purchases.

To estimate the probability that the group who does not use the Simmons card used the coupon for a purchase, we substitute $x_1 = 2$ and $x_2 = 0$ into the estimated logistic regression equation.

$$\hat{y} = \frac{e^{-2.146+0.342(2)+1.099(0)}}{1+e^{-2.146+0.342(2)+1.099(0)}} = \frac{e^{-1.462}}{1+e^{-1.462}} = \frac{0.2318}{1.2318} = 0.1882$$

We repeat the same calculation for the group of customers who uses the Simmons credit card for a purchase by substituting $x_1 = 2$ and $x_2 = 1$ into the estimated logistic regression equation.

$$\hat{y} = \frac{e^{-2.146+0.342(2)+1.099(1)}}{1+e^{-2.146+0.342(2)+1.099(1)}} = \frac{e^{-1.462}}{1+e^{-1.462}} = \frac{0.6956}{1.6956} = 0.4102$$

The probability of using the coupon is much higher for customers with a Simmons credit card.

Simmons Stores Example Significance Tests

When testing for the overall significance of the model, we set the hypotheses as

H_0: $\beta_1 = \beta_2 = 0$

H_a: At least one of the two parameters is not equal to zero

The computer output yields a value of $X^2 = 13.63$ and $p-\text{value} = 0.0011$ for the whole model.

Thus, at $\alpha = 0.01$, we reject H_0 and conclude that the overall model is significant.

The hypotheses for the tests of individual significance of the independent variables are

H_0: $\beta_1 = 0$
H_a: $\beta_1 \neq 0$

and

H_0: $\beta_2 = 0$
H_a: $\beta_2 \neq 0$

The computer output provides a value of $X^2 = 7.56$ and $p-\text{value} = 0.0060$ for the variable Spending (x_1), and a value of $X^2 = 6.41$ and $p-\text{value} = 0.0013$ for the variable Card (x_2.)

Thus, at $\alpha = 0.01$, we reject both H_0: $\beta_1 = 0$ and H_0: $\beta_2 = 0$ and conclude that both independent variables are statistically significant.

Managerial Use of Logistic Regression Results

We can expand on the previous calculations and estimate the probabilities for values of annual spending ranging from $1,000 to $7,000 both for customers who have a Simmons credit card and customers who do not have a Simmons credit card.

	Annual Spending						
Credit Card	**$1,000**	**$2,000**	**$3,000**	**$4,000**	**$5,000**	**$6,000**	**$7,000**
Yes	0.3307	0.4102	0.4948	0.5796	0.6599	0.7320	0.7936
No	0.1414	0.1881	0.246	0.3148	0.3927	0.4765	0.5617

Suppose Simmons wants to send the promotional catalog only to customers who have a 0.40 or higher probability of using the coupon.

Using the estimated probabilities in the table, Simmons promotion strategy would be to only send the catalog to those customers who

- spent $2,000 or more annually, if they have the Simmons credit card.
- spent $6,000 or more annually, if they do not have the Simmons credit card.

The Odds Ratio in Logistic Regression

The odds in favor of an event occurring is defined as the probability the event will occur divided by the probability the event will not occur.

Because in logistic regression the event of interest is always $y = 1$, given a particular set of values for the independent variables, the odds in favor of $y = 1$ can be calculated as

$$\textbf{odds} = \frac{\boldsymbol{P(y = 1|x_1, x_2, \ldots, x_p)}}{\boldsymbol{P(y = 0|x_1, x_2, \ldots, x_p)}} = \frac{\boldsymbol{P(y = 1|x_1, x_2, \ldots, x_p)}}{\boldsymbol{1 - P(y = 1|x_1, x_2, \ldots, x_p)}}$$

The **odds ratio** measures the impact on the odds of a one-unit increase in only one of the independent variables.

$$\textbf{odds ratio} = \frac{\textbf{odds}_1}{\textbf{odds}_0}$$

Where, odds_1 is the odds that $y = 1$, given an increase of one unit in the independent variable
odds_0 is the odds that $y = 1$, given no change in the value of the independent variable

Interpreting the Odds for the Simmons Credit Card

Suppose we want to compare the odds of using the coupon for customers who spend \$2,000 annually and have a Simmons credit card ($x_1 = 2$ and $x_2 = 1$) to the odds of those who spend \$2,000 annually and do not have a Simmons credit card ($x_1 = 2$ and $x_2 = 0$).

Because we are interested in interpreting the effect of a one-unit increase in x_2, we have

$$\text{odds}_1 = \frac{P(y = 1 | x_1 = 2 \text{ and } x_2 = 1)}{1 - P(y = 1 | x_1 = 2 \text{ and } x_2 = 1)} \quad \text{and} \quad \text{odds}_0 = \frac{P(y = 1 | x_1 = 2 \text{ and } x_2 = 0)}{1 - P(y = 1 | x_1 = 2 \text{ and } x_2 = 0)}$$

We have already calculated probabilities for $x_2 = 1$ as 0.4102, and for $x_2 = 0$ as 0.1882.

$$\text{estimate of odds}_1 = \frac{0.4102}{1 - 0.4102} = 0.6956 \quad \text{and} \quad \text{estimate of odds}_0 = \frac{0.1882}{1 - 0.1882} = 0.2318$$

Hence, the estimated odds ratio is $\text{odds ratio} = \dfrac{\text{odds}_1}{\text{odds}_0} = \dfrac{0.6956}{0.2318} = 3.000$

We conclude that the estimated odds in favor of using the coupon are three times greater for customers who have a Simmons credit card than for those who do not.

Interpreting the Odds for Spending at Simmons

The value for Spending in the Odds Ratio column of the logistic regression computer output tells us that the estimated odds in favor of using the coupon is 1.4073 times greater for a one-unit increase in the independent variable Spending.

To address the change in the odds for an increase of more than one unit for an independent variable, we take advantage of the following relationship.

$$\textbf{odds ratio} = e^{c\beta_i}$$ where c is the number of unit increase in the independent variable i

For example, in the Simmons Stores example, for $c = 1$, we have

$$e^{b_1} = e^{0.342} = 1.407 \quad \text{and} \quad e^{b_2} = e^{1.099} = 3.000$$

Suppose that we want to compare the odds of using the coupon for customers who spend \$5,000 annually ($x_1 = 5$) to the odds for those who spend \$2,000 annually ($x_1 = 2$).

In this case $c = 5 - 2 = 3$, and the corresponding estimated odds ratio is

$$\text{estimate odd ratio for a \$3,000 spending increase} = e^{c\beta_1} = e^{3(0.342)} = e^{1.026} = 2.79$$

Confidence Interval for the Odds Ratio

In general, the odds ratio enables us to compare the odds for two different events:

- If the value of an odds ratio is 1, the odds for both events are the same.
- On the other hand, if the value of an odds ratio is greater than 1, the independent variable has a positive impact on the probability of the event occurring.

The Odds Ratio table in the computer output for the Simmons Stores example also provides a 95% confidence interval for each of the odds ratios.

For example, the point estimate of the odds ratio for x_1 is 1.4073 and the 95% confidence interval is 1.0936 to 1.8109.

Because the confidence interval does not contain the value of 1, we can conclude that x_1 has a significant relationship with the estimated odds ratio.

Similarly, because the 95% confidence interval for the odds ratio for x_2 does not contain the value of 1 either, we can conclude that x_2 also has a significant relationship with the odds ratio.

Logit Transformation

The **logit** of a logistic regression equation, denoted $g(x_1, x_2, \ldots, x_p)$, is the natural logarithm of the odds in favor of $y = 1$, and a linear function of the independent variables.

$$\boldsymbol{g(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}$$

With the logit, we can write the logistic regression equation as $E(y) = \dfrac{e^{g(x_1, x_2, \ldots, x_p)}}{1 + e^{g(x_1, x_2, \ldots, x_p)}}$

Once we estimate the parameters in the logistic regression equation, we can compute the **estimated logit**, denoted $\hat{g}(x_1, x_2, \ldots, x_p)$ as

$$\hat{g}(x_1, x_2, \ldots, x_p) = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

For the Simmons Stores example, we have $\hat{g}(x_1, x_2) = -2.146 + 0.342x_1 + 1.099x_2$

And the estimated logistic regression equation is $\hat{y} = \dfrac{e^{\hat{g}(x_1, x_2)}}{1 + e^{\hat{g}(x_1, x_2)}} = \dfrac{e^{-2.146 + 0.342x_1 + 1.099x_2}}{1 + e^{-2.146 + 0.342x_1 + 1.099x_2}}$

Practical Advice: Big Data and Hypothesis Testing in Multiple Regression

In multiple regression, as the sample size increases,

- the *p*-value for the F test, used to determine whether a significant relationship exists between the dependent variable and the set of all independent variables, decreases.
- the *p*-value for each of *t* test used to determine whether a significant relationship exists between the dependent variable and an individual independent variable, decreases.
- the confidence interval for the slope parameter associated with each individual independent variable narrows.
- the confidence interval for the mean value of y narrows.
- the prediction interval for an individual value of y narrows.

Besides the effects due to sample size and nonsampling error already introduced for simple linear regression, multicollinearity may also cause the estimated slope coefficients in multiple regression to be misleading.