統計學 (二)

Anderson's Statistics for Business & Economics (14/E)

Chapter 15: Multiple Regression

上課時間地點: 二 D56, 資訊 140306
授課教師: 吳漢銘 (國立政治大學統計學系副教授)
教學網站: http://www.hmwu.idv.tw

系級: _____ 學號: _____ 姓名: _____

15.1 Multiple Regression Model

1. (Recall) that the variable being predicted or explained is called the _____ variable and the variable being used to predict or explain the dependent variable is called the _____ variable.

2. Multiple regression analysis is the study of how a dependent variable y is related to _____ variables. In the general case, we will use _____ to denote the number of independent variables.

3. The concepts of a regression model and a regression equation introduced in the preceding chapter are _____ in the multiple regression case.

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

$$\text{_____} \qquad (15.1)$$

5. In the multiple regression model, $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ are the _____ and the error term ϵ is a _____. y is a linear function of $x_1, x_2, \cdots, x_p$ plus the error term ϵ.

6. The error term accounts for the _____ in y that _____ by the linear effect of the p independent variables.

7. (**Multiple regression equation**):The equation that describes how the mean value of y is related to $x_1, x_2, \cdots, x_p$ is called the multiple regression equation.

$$\underline{\qquad\qquad\qquad\qquad\qquad\qquad} \quad (15.2)$$

under the assumption that the mean or expected value of ϵ is zero.

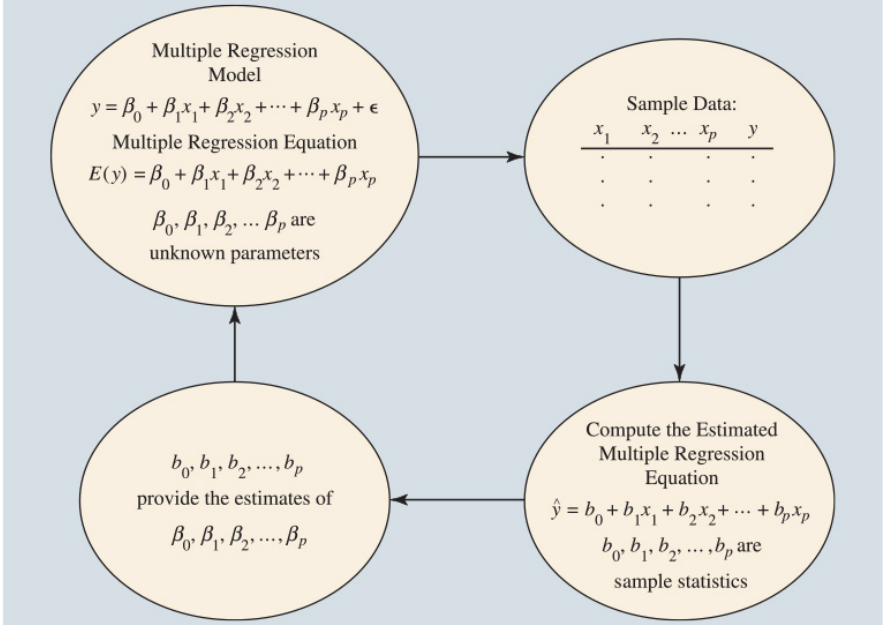
8. **The estimated multiple regression equation**:

$$\underline{\qquad\qquad\qquad\qquad\qquad\qquad} \quad (15.3)$$

where $b_0, b_1, b_2, \cdots, b_p$ are the estimates of $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ and $\hat{y}$ is the predicted value of the dependent variable

9. (Figure 15.1)

FIGURE 15.1 The Estimation Process for Multiple Regression



15.2 Least Squares Method

1. The least squares method is used to develop the estimated multiple regression equation:

   $$\underline{\qquad\qquad\qquad\qquad} \tag{15.4}$$

   where y_i is observed value of the dependent variable for the ith observation, $\hat{y}_i$ is predicted value of the dependent variable for the ith observation

2. In multiple regression, however, the presentation of the formulas for the regression coefficients $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ involves the use of _____ and is beyond the scope of this text.

3. Therefore, in presenting multiple regression, we focus on how statistical software can be used to obtain the estimated regression equation and other information. The emphasis will be on how to _____ the computer output rather than on how to make the multiple regression computations.

An Example: Butler Trucking Company

1. The Butler Trucking Company, an independent trucking company in southern California.

2. A major portion of Butler's business involves deliveries throughout its local area. To develop better work schedules, the managers want to predict the total daily travel time for their drivers.

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

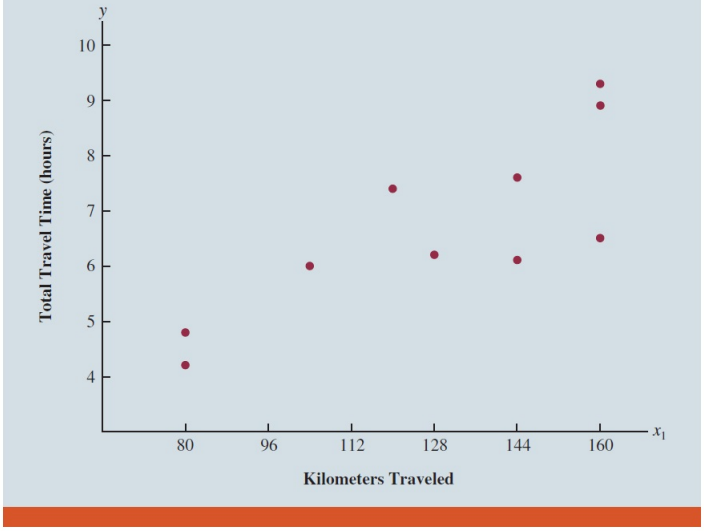
   (b) (Table 15.1)(Figure 15.2) A simple random sample of 10 driving assignments provided the data shown in Table 15.1 and the scatter diagram shown in Figure 15.2.

TABLE 15.1 Preliminary Data for Butler Trucking

Driving Assignment	x_1 = Kilometers Traveled	y = Travel Time (hours)
1	160	9.3
2	80	4.8
3	160	8.9
4	160	6.5
5	80	4.2
6	128	6.2
7	120	7.4
8	104	6.0
9	144	7.6
10	144	6.1

Source: PC Magazine website, April, 2015. (https://www.pcmag.com/reviews/monitors)

FIGURE 15.2 Scatter Diagram of Preliminary Data for Butler Trucking



(c) After reviewing this scatter diagram, the managers hypothesized that the simple linear regression model $y = \beta_0 + \beta_1 x_1 + \epsilon$ could be used to describe the relationship between the total travel time (y) and the number of miles traveled (x_1).

(d) (Figure 15.3) we show statistical software output from applying simple linear regression to the data in Table 15.1. The estimated regression equation is

_____

i. At the 0.05 level of significance, the F value of _____ and its corresponding p-value of _____ indicate that the relationship is significant; that is, we can reject $H_0 : \beta_1 = 0$ because the p-value is less than $\alpha = 0.05$.

ii. Note that the same conclusion is obtained from the t value of _____ and its associated p-value of _____.

iii. Thus, we can conclude that the relationship between the total travel time and the number of miles traveled is _____; longer travel times are associated with more miles traveled.

FIGURE 15.3 Output for Butler Trucking with One Independent Variable

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	15.871	15.8713	15.81	.004
Error	8	8.029	1.0036		
Total	9	23.900			

Model Summary

S	R-sq	R-sq (adj)
1.00179	66.41%	62.21%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	1.27	1.40	.91	.390
Kilometers	.0424	.0107	3.98	.004

Regression Equation

Time = 1.27 + .0424 Kilometers

iv. With a coefficient of determination (expressed as a percentage) of _____, we see that _____ in travel time can be explained by the linear effect of the number of miles traveled.

3. (Table 15.2) The managers might want to consider adding a second independent variable (number of deliveries) to explain some of the remaining variability in the dependent variable.

TABLE 15.2 Data for Butler Trucking with Kilometers Traveled (x_1) and Number of Deliveries (x_2) as the Independent Variables

Driving Assignment	x_1 = Kilometers Traveled	x_2 = Number of Deliveries	y = Travel Time (hours)
1	160	4	9.3
2	80	3	4.8
3	160	4	8.9
4	160	2	6.5
5	80	2	4.2
6	128	2	6.2
7	120	3	7.4
8	104	4	6.0
9	144	3	7.6
10	144	2	6.1

4. (Figure 15.4) Computer output with both miles traveled (x_1) and number of deliveries (x_2) as independent variables is shown in Figure 15.4. The estimated regression equation is

$$\hat{y} = \_\_\_\_\_ \quad (15.6)$$

FIGURE 15.4 Output for Butler Trucking with Two Independent Variables

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	21.6006	10.8003	32.88	.000
Error	7	2.2994	.3285		
Total	9	23.900			

Model Summary

S	R-sq	R-sq (adj)
.573142	90.38%	87.63%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	−.869	.952	−.91	.392
Kilometers	.03821	.00618	6.18	.000
Deliveries	.923	.221	4.18	.004

Regression Equation

Time = −.869 + .03821 Kilometers + 0.923 Deliveries

Note on Interpretation of Coefficients

1. One observation can be made at this point about the relationship between the estimated regression equation with only the miles traveled as an independent variable and the equation that includes the _____ as a second independent variable.

2. The value of _____ is not the same in both cases. In simple linear regression, we interpret β_1 as an estimate of the change in y for a _____ in the independent variable.

3. In multiple regression analysis, we interpret each regression coefficient as follows: b_i represents an estimate of the _____ corresponding to a _____ _____ when all other independent variables are _____.

4. Butler Trucking example

   (a) $\beta_1 = 0.06113$, 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 is 0.06113 hours.

   (b) $\beta_2 = 0.923$, 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 is 0.923 hours.

😍 **EXERCISES 15.2**: 1, 5, 6

15.3 Multiple Coefficient of Determination

1. In simple linear regression, we showed that the total sum of squares can be partitioned into two components: the sum of squares due to regression and the sum of squares due to error. The same procedure applies to the sum of squares in multiple regression.

   _____ (15.7)

   where

   SST: total sum of squares = _____.

   SSR: sum of squares due to regression = _____.

   SSE: sum of squares due to error = _____.

2. Example Butler Trucking problem (Figure 15.4)

FIGURE 15.4 Output for Butler Trucking with Two Independent Variables

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	21.6006	10.8003	32.88	.000
Error	7	2.2994	.3285		
Total	9	23.900			

Model Summary

S	R-sq	R-sq (adj)
.573142	90.38%	87.63%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	−.869	.952	−.91	.392
Kilometers	.03821	.00618	6.18	.000
Deliveries	.923	.221	4.18	.004

Regression Equation

Time = −.869 + .03821 Kilometers + 0.923 Deliveries

$SST = 23.900$, $SSR = 21.6006$, and $SSE = 2.2994$.

3. With only one independent variable (number of miles traveled), the output in Figure 15.3 shows that $SST = 23.900$, $SSR = 15.871$, and $SSE = 8.029$. The value of SST is the same in both cases because it does not depend on $\hat{y}$, but SSR increases and SSE decreases when a second independent variable (number of deliveries) is added.

4. The multiple coefficient of determination, denoted R^2, measures the goodness of fit for the estimated multiple regression equation.

(15.8)

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

6. Hence, when multiplied by 100, it can be interpreted as the percentage of the variability in y that can be explained _____.

7. Example In the two-independent-variable Butler Trucking example, with $SSR = 21.6006$ and $SST = 23.900$, we have $R^2 = 21.6006/23.900 = 0.9038$.

8. Therefore, 90.38% 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.

9. (Figure 15.3) the R-sq value for the estimated regression equation with only one independent variable, number of miles traveled (x_1), is 66.41%. Thus, the percentage of the variability in travel times that is explained by the estimated regression equation increases from _____ when number of deliveries is added as a second independent variable.

10. In general, R^2 always increases as independent variables are added to the model.

11. Many analysts prefer adjusting R^2 for the number of independent variables to avoid _____ the impact of adding an independent variable on the amount of variability exlained by the estimated regression equation.

12. With n denoting the number of observations and p denoting the number of independent variables, the adjusted multiple coefficient of determination is computed as follows:

(15.9)

13. Example With $n = 10$ and $p = 2$, we have

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

14. Thus, after adjusting for the two independent variables, we have an adjusted multiple coefficient of determination of 0.8763. This value (expressed as a percentage) is provided in the output in Figure 15.4 as _____.

15. If the value of R^2 is small and the model contains a large number of independent variables, the adjusted coefficient of determination can take a _____; in such cases, statistical software usually sets the adjusted coefficient of determination to _____.

😍 **EXERCISES 15.3**: 11, 14, 15

15.4 Model Assumptions

1. The multiple regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \qquad (15.10)$$

2. The assumptions about the _____ in the multiple regression model:

   (1) The error term ϵ is a random variable with mean or expected value of zero; that is, _____.

   Implication: For given values of $x_1, x_2, \cdots, x_p$, the expected, or average, value of y is given by

$$E(y) = \_\_\_\_\_ \qquad (15.11)$$

   Equation (15.11) is the _____. $E(y)$ represents the average of all possible values of y that might occur for the given values of $x_1, x_2, \cdots, x_p$.

   (2) The variance of ϵ is denoted by σ^2 and is the same for all values of the independent variables $x_1, x_2, \cdots, x_p$; that is, _____.

   Implication: The variance of y about the regression line equals _____ and is the same for all values of $x_1, x_2, \cdots, x_p$.

   (3) The values of ϵ are _____.

   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.

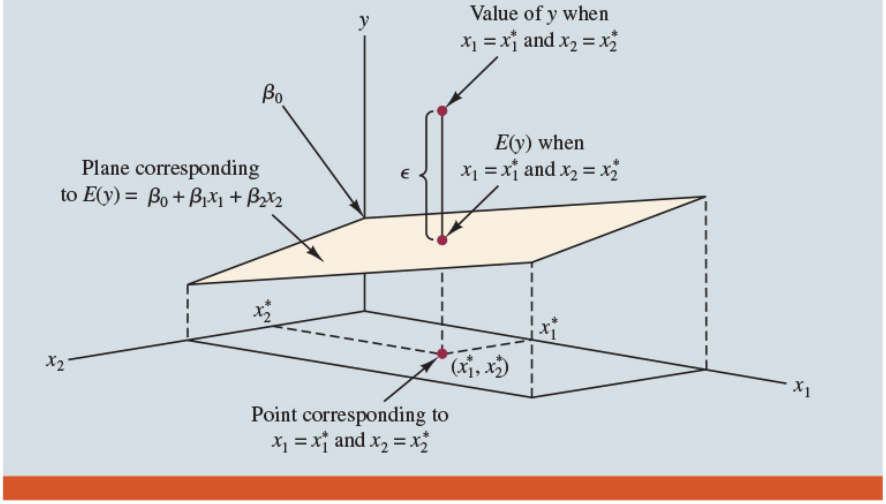
   (4) The error term ϵ is a _____ random variable reflecting the deviation between the _____ value and the _____ given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.

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

3. (Figure 15.5) Consider the following two-independent-variable multiple regression equation.

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

FIGURE 15.5 Graph of the Regression Equation for Multiple Regression Analysis with Two Independent Variables



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

5. In regression analysis, the term response variable is often used in place of the term _____. Furthermore, since the multiple regression equation generates a plane or surface, its graph is called a _____.

15.5 Testing for Significance

1. In simple linear regression, both _____ and an _____ provide the same conclusion; that is, if the null hypothesis is rejected, we conclude that _____.

2. In multiple regression, the t test and the F test have different purposes.

   (a) The F test is used to determine whether a significant relationship exists between the dependent variable and the set of _____ the independent variables; we will refer to the F test as the test for _____.

   (b) If the F test shows an overall significance, the _____ 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 _____.

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

F Test

1. The hypotheses for the F test involve the parameters of the multiple regression model.

   H_0 : _____

   H_a : One or more of the parameters are not equal to zero

2. If H_0 is rejected, the test gives us _____ to conclude that one or more of the parameters are not equal to zero and that the _____ between y and the set of independent variables $x_1, x_2, \cdots, x_p$ is _____.

3. However, if H_0 cannot be rejected, we do not have _____ to conclude that a significant relationship is present.

4. (Review)(Chapter 14)

   (a) A mean square is a _____ divided by its corresponding degrees of freedom.

   (b) In the multiple regression case, the total sum of squares (SST) has _____ degrees of freedom, the sum of squares due to regression (SSR) has _____ degrees of freedom, and the sum of squares due to error (SSE) has _____ degrees of freedom.

(c) Hence, the mean square due to regression (MSR) is _____ and the mean square due to error (MSE) is _____.

(d) MSE provides an unbiased estimate of _____, the variance of the error term ϵ.

(e) If _____ is true, _____ also provides an unbiased estimate of σ^2, and the value of MSR/MSE should be close to _____.

(f) However, if H_0 is false, MSR _____ σ^2 and the value of MSR/MSE becomes _____.

5. To determine how large the value of _____ must be to reject H_0, we make use of the fact that if _____ and the _____ about the multiple regression model are _____, the sampling distribution of MSR/MSE is an _____ distribution with _____ degrees of freedom in the numerator and _____ in the denominator.

6. **F test for overall significance**

(a) Hypothesis:

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

$$H_a \ : \ \text{One or more of the parameters are not equal to zero}$$

(b) Test statistic:

_____ (15.14)

(c) Rejection rule:

i. p-value approach: Reject H_0 if _____.

ii. Critical value approach: Reject H_0 if _____

TABLE 15.3 ANOVA Table for a Multiple Regression Model with p Independent Variables

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

7. Example Butler Trucking Company

(a) Hypotheses:

$$H_0 \ : \ \beta_1 = \beta_2 = 0$$
$$H_a \ : \ \beta_1 \text{ and/or } \beta_2 \text{ is not equal to zero}$$

(b) (Figure 15.6)

FIGURE 15.6 Output for Butler Trucking with Two Independent Variables, Kilometers Traveled (x_1) and Number of Deliveries (x_2)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	21.6006	10.8003	32.88	.000
Error	7	2.2994	.3285		
Total	9	23.900			

Model Summary

S	R-sq	R-sq (adj)
.573142	90.38%	87.63%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	−.869	.952	−.91	.392
Kilometers	.03821	.00618	6.18	.000
Deliveries	.923	.221	4.18	.004

Regression Equation

Time = −.869 + .03821 Kilometers + .923 Deliveries

(c) $MSR = 10.8003$ and $MSE = 0.3285$, $F =$ _____. Using $\alpha = 0.01$, _____. With $F = 32.88 > 9.55$, we reject $H_0 : \beta_1 = \beta_2 = 0$.

(d) Using $\alpha = 0.01$, the p-value $= 0.000$ indicates that we can reject $H_0 : \beta_1 = \beta_2 = 0$ because the p-value is less than $\alpha = 0.01$.

(e) Conclude that a _____ is present between travel time y and the two independent variables, miles traveled and number of deliveries.

t Test

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

2. **The t test for individual significance**

   (a) Hypothesis: For any parameter β_i

   $$\begin{aligned} H_0 &: \_\_\_\_\_ \\ H_a &: \beta_i \neq 0 \end{aligned}$$

   (b) Test statistic:

   $$\_\_\_\_\_ \qquad (15.15)$$

   (c) Rejection rule:

   i. p-value approach: Reject H_0 if p-value $\leq \alpha$.

   ii. Critical value approach: Reject H_0 if _____ or if _____.

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

   補充:

   The multiple regression model

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \epsilon,$$

   or

   $$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \quad i = 1, \cdots, n.$$

   (a) In the matrix notation:

   $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \qquad \text{or} \qquad \mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}.$$

   (b) Use Least-squares to fit a regression line to the data $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,p-1}\}$

   $$Q(\boldsymbol{\beta}) \;=\; (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2.$$

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0} \\ \Rightarrow \quad & (\mathbf{X}'\mathbf{X})\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \\ \Rightarrow \quad & \hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \end{aligned}$$

(c) Variance of the sampling distribution of $b_i, i = 1, 2, \cdots, p$.

$$Var(b_i) = \frac{\sigma^2}{(n-1)S_{x_i}^2(1 - R_i^2)},$$

where $S_{x_i}^2$ is the sample variance of variable x_i and R_i^2 is R-square of the regression of x_i on the rest of the explanatory variables of the models (including the constant term). Note that the variance should be conditional on the observed values of the explanatory variables.

4. Example Butler Trucking Company

(a) (Figure 15.6) that shows the output for the t-ratio calculations:

$$b_1 = 0.06113, b_2 = 0.923, s_{b_1} = 0.00989, s_{b_2} = 0.221$$

(b) The test statistic for the hypotheses involving parameters β_1 and β_2:

$$t = 0.06113/0.00989 = 6.18, \quad t = 0.923/0.221 = 4.18$$

(c) Using $\alpha = 0.01$, the p-values of _____ and _____ in the output indicate that we can reject $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. Hence, both parameters are statistically significant.

(d) Alternatively, _____. With $6.18 > 3.499$, we reject $H_0 : \beta_1 = 0$. Similarly, with $4.18 > 3.499$, we reject $H_0 : \beta_2 = 0$.

Multicollinearity

1. We use the term _____ in regression analysis to refer to any variable being used to predict or explain the value of the dependent variable.

2. The term does not mean, however, that the independent variables _____ are independent in any statistical sense. On the contrary, most independent variables in a multiple regression problem are _____ to some degree with one another.

3. Example Butler Trucking Example

   (a) Butler Trucking example involves the two independent variables x_1 (miles traveled) and x_2 (number of deliveries), we could treat the miles traveled as the dependent variable and the number of deliveries as the independent variable to determine whether those two variables are themselves related.

   (b) Compute the sample correlation coefficient $r(x_1, x_2) = 0.16$ and find that some degree of linear association between the two independent variables.

4. In multiple regression analysis, _____ refers to the correlation among the independent variables.

5. Example Modified Butler Trucking Example, the potential problems of multicollinearity.

   (a) Consider a modification of the Butler Trucking example. Instead of x_2 being the number of deliveries, let x_2 denote the number of gallons of gasoline consumed. Clearly, x_1 (the miles traveled) and x_2 are related; that is, we know that the number of gallons of gasoline used depends on the number of miles traveled.

   (b) We would conclude logically that x_1 and x_2 are highly correlated independent variables.

   (c) Assume that we obtain the equation $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ and find that the F test shows the relationship to be significant. Then suppose we conduct a t test on β_1 to determine whether $\beta_1 \neq 0$, and we cannot reject $H_0 : \beta_1 = 0$. Does this result mean that travel time is not related to miles traveled? Not necessarily.

   (d) What it probably means is that with _____, x_1 does not make a significant contribution to determining the value of y.

(e) This interpretation makes sense in our example; if we know the amount of gasoline consumed (x_2), we do not gain much additional information useful in predicting y by knowing the miles traveled (x_1).

(f) Similarly, a t test might lead us to conclude $\beta_2 = 0$ on the grounds that, with x_1 in the model, knowledge of the amount of gasoline consumed does not add much.

6. To summarize, in _____ for the significance of individual parameters, the difficulty caused by multicollinearity is that it is possible to conclude that _____ of the individual parameters is significantly different from zero when an _____ on the _____ multiple regression equation indicates a significant relationship.

7. Statisticians have developed several _____ for determining whether multicollinearity is high enough to cause problems.

8. According to the rule of thumb test, multicollinearity is a potential problem if the absolute value of the _____ exceeds _____ for any two of the independent variables.

9. The other types of tests are more advanced and beyond the scope of this text. If possible, every attempt should be made to avoid including independent variables that are highly correlated.

10. When multicollinearity is severe,

(a) it is not possible to determine the separate effect of any particular independent variable on the dependent variable.

(b) we can have difficulty interpreting the results of t tests on the individual parameters.

(c) Least squares estimates may have the wrong sign.

11. 補充:

(a) Multicollinearity in Regression Analysis: Problems, Detection, and Solutions

`https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/`

(b) Multicollinearity in Regression: Why it is a problem? How to check and fix it

https://towardsdatascience.com/multi-collinearity-in-regression-fe7a2c1467ea

(c) Eight Ways to Detect Multicollinearity

https://www.theanalysisfactor.com/eight-ways-to-detect-multicollinearity/

(d) Multicollinearity (Wikipedia)

https://en.wikipedia.org/wiki/Multicollinearity

😍 **EXERCISES 15.5**: 19, 23, 24

15.6 Using the Estimated Regression Equation for Estimation and Prediction

1. The procedures for estimating the mean value of y and predicting an individual value of y in multiple regression are similar to those in regression analysis involving one independent variable.

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

3. Example Butler Trucking example

   (a) We want to use the estimated regression equation involving x_1 (miles traveled) and x_2 (number of deliveries) to develop two interval estimates:

      i. A _____ of the mean travel time for all trucks that travel 100 miles and make two deliveries.

      ii. A _____ of the travel time for one specific truck that travels 100 miles and makes two deliveries

   (b) Using the estimated regression equation $\hat{y} = -0.869 + 0.06113x_1 + 0.923x_2$ with $x_1 = 100$ and $x_2 = 2$, we obtain

$$\hat{y} = \_\_\_\_\_$$

Hence, the point estimate of travel time in both cases is approximately seven hours.

(c) To develop interval estimates for the mean value of y and for an individual value of y, we use a procedure similar to that for regression analysis involving one independent variable. The formulas required are beyond the scope of the text, but statistical _____ for multiple regression analysis will often provide confidence intervals once the values of $x_1, x_2, \cdots, x_p$ are specified by the user.

(d) (Table 15.4)

TABLE 15.4 The 95% Confidence and Prediction Intervals for Butler Trucking

Value of x_1	Value of x_2	95% Confidence Interval		95% Prediction Interval	
		Lower Limit	Upper Limit	Lower Limit	Upper Limit
160	4	8.135	9.742	7.363	10.514
80	3	4.127	5.789	3.369	6.548
160	4	8.135	9.742	7.363	10.514
160	2	6.258	7.925	5.500	8.683
80	2	3.146	4.924	2.414	5.656
128	2	5.232	6.505	4.372	7.366
120	3	6.037	6.936	5.059	7.915
104	4	5.960	7.637	5.205	8.392
144	3	6.917	7.891	5.964	8.844
144	2	5.776	7.184	4.953	8.007
120	4	6.669	8.152	5.865	8.955

(e) Note that the interval estimate for an individual value of y is _____ the interval estimate for the expected value of y. This difference simply reflects the fact that for given values of x_1 and x_2 we can estimate the mean travel time for all trucks with _____ than we can predict the travel time for one specific truck.

😍 **EXERCISES 15.6**: 27, 29

15.7 Categorical Independent Variables

(a) Thus far, the examples we have considered involved _____ independent variables such as student population, distance traveled, and number of deliveries.

(b) In many situations, however, we must work with _____ independent variables such as gender (male, female), method of payment (cash, credit card, check), and so on.

An Example: Johnson Filtration, Inc.

(a) (Background) Johnson Filtration, Inc., provides maintenance service for water-filtration systems throughout southern Florida. Customers contact Johnson with requests for maintenance service on their water-filtration systems. To estimate the service time and the service cost, Johnson′ s managers want to predict the repair time necessary for each maintenance request.

(b) (Dependent variable/Independent variables) Hence, repair time in hours is the dependent variable. Repair time is believed to be related to two factors, the number of months since the last maintenance service and the type of repair problem (mechanical or electrical).

(c) (Data)(Table 15.5)

TABLE 15.5 Data for the Johnson Filtration Example

Service Call	Months Since Last Service	Type of Repair	Repair Time in Hours
1	2	Electrical	2.9
2	6	Mechanical	3.0
3	8	Electrical	4.8
4	3	Mechanical	1.8
5	2	Electrical	2.9
6	7	Electrical	4.9
7	9	Mechanical	4.2
8	8	Mechanical	4.8
9	4	Electrical	4.4
10	6	Electrical	4.5

(d) (SLR) Let y denote the repair time in hours and x_1 denote the number of months since the last maintenance service. The regression model that uses only x_1 to predict y is $y = \beta_0 + \beta_1 x_1 + \epsilon$

(e) (Figure 15.7)

FIGURE 15.7 Output for Johnson Filtration with Months Since Last Service (x_1) as the Independent Variable

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	5.596	5.5960	9.17	.016
Error	8	4.880	.6100		
Total	9	10.476			

Model Summary

S	R-sq	R-sq (adj)
.781022	53.42%	47.59%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	2.147	.605	3.55	.008
Months Since Last Service	.304	.100	3.03	.016

Regression Equation

Repair Time (hours) = 2.147 + .304 Months Since Last Service

i. The estimated regression equation is _____.

ii. At the 0.05 level of significance, the p-value of _____ for the t (or F) test indicates that the number of months since the last service is significantly related to repair time.

iii. R-sq = _____ indicates that x_1 alone explains _____ of the _____ in repair time.

4. To incorporate the type of repair into the regression model, we define

$$x_2 = \begin{cases} \_\_\_\_\_, & \text{if the type of repair is mechanical} \\ \_\_\_\_\_, & \text{if the type of repair is electrical} \end{cases}$$

5. In regression analysis x_2 is called a _____ or _____.

6. Using this dummy variable, we can write the multiple regression model as

$$y = \_\_\_\_\_$$

7. (Table 15.6) Data for the Johnson Filtration Example with Type of Repair Indicated by a Dummy Variable ($x_2 = 0$ for Mechanical; $x_2 = 1$ for Electrical)

TABLE 15.6 Data for the Johnson Filtration Example with Type of Repair Indicated by a Dummy Variable ($x_2 = 0$ for Mechanical; $x_2 = 1$ for Electrical)

Customer	Months Since Last Service (x_1)	Type of Repair (x_2)	Repair Time in Hours (y)
1	2	1	2.9
2	6	0	3.0
3	8	1	4.8
4	3	0	1.8
5	2	1	2.9
6	7	1	4.9
7	9	0	4.2
8	8	0	4.8
9	4	1	4.4
10	6	1	4.5

8. (Figure 15.7) Output for Johnson Filtration with Months Since Last Service (x_1) as the Independent Variable

FIGURE 15.8 Output for Johnson Filtration with Months Since Last Service (x_1) and Type of Repair (x_2) as the Independent Variables

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	9.0009	4.50046	21.36	.001
Error	7	1.4751	.21073		
Total	9	10.4760			

Model Summary

S	R-sq	R-sq (adj)
.459048	85.92%	81.90%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	.930	.467	1.99	.087
Months Since Last Service	.3876	.0626	6.20	.000
Type of Repair	1.263	.314	4.02	.005

Regression Equation

Repair Time (hours) = .930 + .3876 Months Since Last Service + 1.263 Type of Repair

(a) The estimated multiple regression equation is

_____ (15.17)

(b) At the 0.05 level of significance, the p-value of _____ associated with the F test (_____) indicates that the regression relationship is significant.

(c) The t test shows that both months since last service (p-value = _____) and type of repair (p-value = _____) are statistically significant.

(d) In addition, R-Sq = _____ and R-Sq (adj) = _____ indicate that the estimated regression equation does a good job of explaining the variability in repair times.

(e) Thus, equation (15.17) should prove helpful in predicting the repair time necessary for the various service calls.

Interpreting the Parameters

1. The multiple regression equation for the Johnson Filtration example is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (15.18)$$

2. Consider the case when $x_2 = 0$ (mechanical repair). Using _____ to denote the mean or expected value of repair time given a mechanical repair, we have

$$E(y|\text{mechanical}) = \_\_\_\_\_ = \_\_\_\_\_ \qquad (15.19)$$

3. Similarly, for an electrical repair ($x_2 = 1$), we have

$$E(y|\text{electrical}) = \_\_\_\_\_ = \_\_\_\_\_ \qquad (15.20)$$

4. Comparing equations (15.19) and (15.20), we see that the mean repair time is a linear function of _____ for both mechanical and electrical repairs. The slope of both equations is _____, but the _____ differs.

5. The y-intercept is _____ in equation (15.19) for mechanical repairs and _____ in equation (15.20) for electrical repairs.

6. The interpretation of β_2 is that it indicates the _____ between the _____ for an electrical repair and the mean repair time for a mechanical repair.

   (a) If _____, the mean repair time for an electrical repair will be _____ that for a mechanical repair;

(b) if _____, the mean repair time for an electrical repair will be _____ that for a mechanical repair.

(c) if _____, there is _____ in the mean repair time between electrical and mechanical repairs and the type of repair is _____ to the repair time.

7. Using the estimated multiple regression equation $\hat{y} = 0.93 + 0.3876x_1 + 1.263x_2$, we see that 0.93 is the estimate of β_0 and 1.263 is the estimate of β_2.

8. Thus, when $x_2 = 0$ (mechanical repair)

$$\hat{y} = 0.93 + 0.3876x_1 \qquad (15.21)$$

and when $x_2 = 1$ (electrical repair)

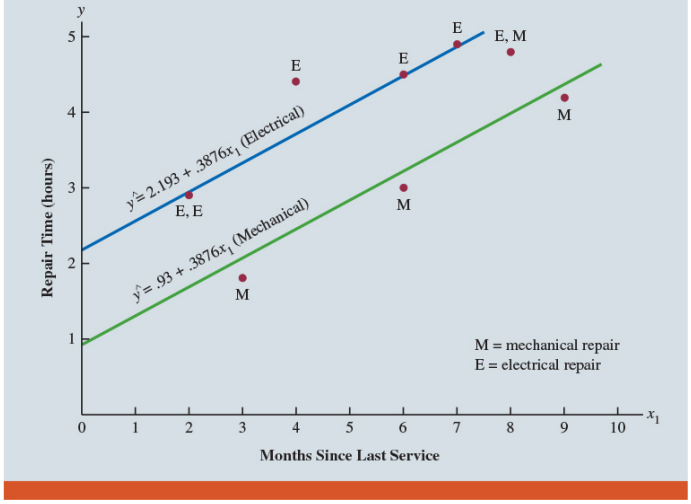
$$\hat{y} = 0.93 + 0.3876x_1 + 1.263(1) = 2.193 + 0.3876x_1 \qquad (15.22)$$

9. In effect, the use of a dummy variable for type of repair provides _____ _____ that can be used to predict the repair time, one corresponding to mechanical repairs and one corresponding to electrical repairs.

10. In addition, with $\beta_2 = 1.263$, we learn that, on average, electrical repairs require _____ than mechanical repairs.

11. (Figure 15.9) Scatter Diagram for the Johnson Filtration Repair Data

FIGURE 15.9 Scatter Diagram for the Johnson Filtration Repair Data from Table 15.6



More Complex Categorical Variables

1. If a categorical variable has k levels, $k-1$ dummy variables are required, with each dummy variable being coded as _____.

2. Example Suppose a manufacturer of copy machines organized the sales territories for a particular state into three regions: A, B, and C. The managers want to use regression analysis to help predict the number of copiers sold per week.

3. With the number of units sold as the dependent variable, they are considering several independent variables (the number of sales personnel, advertising expenditures, and so on).

4. Suppose the managers believe sales region is also an important factor in predicting the number of copiers sold. Because sales region is a categorical variable with three levels, A, B and C, we will need _____ dummy variables to represent the sales region. Each variable can be coded 0 or 1:

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

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

5. We have the following values of x_1 and x_2:

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

6. Observations corresponding to region A would be coded _____; observations corresponding to region B would be coded _____; and observations corresponding to region C would be coded _____.

7. The regression equation relating the expected value of the number of units sold, $E(y)$, to the dummy variables would be written as

$$E(y) = \_\_\_\_\_$$

8. To help us interpret the parameters β_0, β_1, and β_2, consider the following three variations of the regression equation.

$$\begin{aligned} E(y|\text{region A}) &= \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0 \\ E(y|\text{region B}) &= \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1 \\ E(y|\text{region C}) &= \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2 \end{aligned}$$

   (a) Thus, β_0 is the mean or expected value of sales for _____;

   (b) β_1 is the _____ between the mean number of units sold in _____ and the mean number of units sold in _____;

   (c) and β_2 is the _____ between the mean number of units sold in _____ and the mean number of units sold in _____.

9. Two dummy variables were required because sales region is a categorical variable with three levels.

10. The assignment was _____. For example, we could have chosen $x_1 = 1, x_2 = 0$ to indicate region A, $x_1 = 0, x_2 = 0$ to indicate region B, and $x_1 = 0, x_2 = 1$ to indicate region C.

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

    In that case, β_1 would have been interpreted as the mean difference between regions A and B and β_2 as the mean difference between regions C and B.

11. The important point to remember is that when a categorical variable has k levels, $k-1$ dummy variables are required in the multiple regression analysis. Thus, if the sales region example had a fourth region, labeled D, three dummy variables would be necessary. For example, the three dummy variables can be coded as follows.

$$x_1 = \begin{cases} 1, & \text{if sales region B} \\ 0, & \text{otherwise} \end{cases} \quad x_2 = \begin{cases} 1, & \text{if sales region C} \\ 0, & \text{otherwise} \end{cases} \quad x_3 = \begin{cases} 1, & \text{if sales region D} \\ 0, & \text{otherwise} \end{cases}$$

😍 **EXERCISES 15.7**: 32, 34, 35

15.8 Residual Analysis

1. **Standardized Residual for Observation** i

$$\frac{\qquad\qquad}{\qquad\qquad} \tag{15.23}$$

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

2. **Standard Deviation of Residual** i

$$s_{y_i-\hat{y}_i} = \_\_\_\_\_ \tag{15.24}$$

   where s is the standard error of the estimate and h_i is the _____ of observation i.

3. (Chapter 14) the leverage of an observation is determined by how far the values of the _____ are from their _____.

4. The computation of h_i, $s_{y_i-\hat{y}_i}$, and hence the standardized residual for observation i in multiple regression analysis is too complex to be done by hand. However, the standardized residuals can be easily obtained as part of the output from statistical software.

5. Example Butler Trucking example

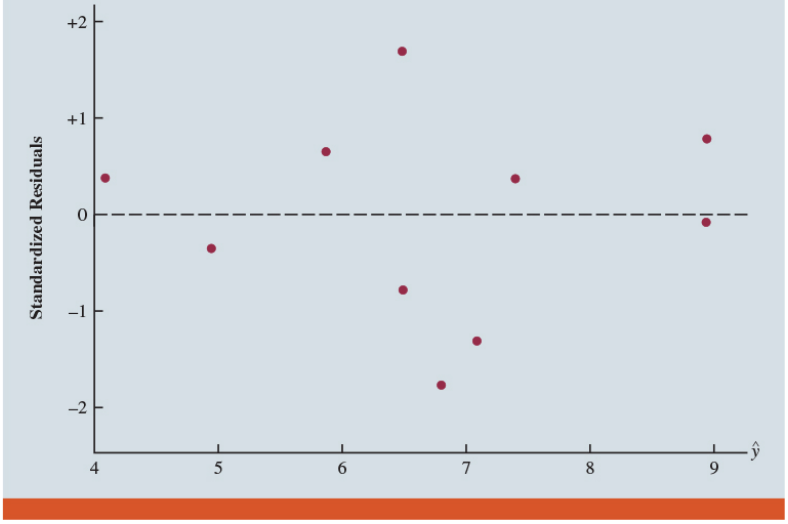
   (a) (Table 15.7) the estimated regression equation $\hat{y} = -0.869 + 0.03821x_1 + 0.923x_2$.

TABLE 15.7 Residuals and Standardized Residuals for the Butler Trucking Regression Analysis

Kilometers Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Predicted Value ($\hat{y}$)	Residual ($y - \hat{y}$)	Standardized Residual
160	4	9.3	8.93846	.361541	.78344
80	3	4.8	4.95830	−.158304	−.34962
160	4	8.9	8.93846	−.038460	−.08334
160	2	6.5	7.09161	−.591609	−1.30929
80	2	4.2	4.03488	.165121	.38167
128	2	6.2	5.86892	.331083	.65431
120	3	7.4	6.48667	.913331	1.68917
104	4	6.0	6.79875	−.798749	−1.77372
144	3	7.6	7.40369	.196311	.36703
144	2	6.1	6.48026	−.380263	−.77639

(b) (Figure 15.10) This standardized residual plot does not indicate any unusual abnormalities. All the standardized residuals are between _____; hence, we have no reason to question the assumption that the error term ϵ is normally distributed. We conclude that the model assumptions are _____.

FIGURE 15.10 Standardized Residual Plot for Butler Trucking



(c) (Recall Section 14.8) A _____ also can be used to determine whether the distribution of ϵ appears to be normal. The same procedure is appropriate for multiple regression.

Detecting Outliers

1. An outlier is an observation that is _____ in comparison with the other data. An outlier does not fit the _____ of the other data.

2. (Chapter 14) An observation is classified as an outlier if the value of its _____ is less than -2 or greater than $+2$.

3. (Table 15.7) Applying this rule to the standardized residuals for the Butler Trucking example, We do not detect any outliers in the data set.

4. In general, the presence of one or more outliers in a data set tends to increase _____, the standard error of the estimate, and hence increase _____, the standard deviation of residual i.

5. Because $s_{y_i-\hat{y}_i}$ appears in the denominator of the formula for the standardized residual (15.23), the size of the standardized residual will _____ as s _____. As a result, even though a residual may be unusually large, the large denominator in expression (15.23) may cause the standardized residual rule to fail to identify the observation as being an outlier.

6. We can circumvent this difficulty by using a form of the standardized residuals called _____.

Studentized Deleted Residuals and Outliers

1. Suppose the ith observation is deleted from the data set and a new estimated regression equation is developed with the remaining $n-1$ observations.

2. Let _____ denote the standard error of the estimate based on the data set with the _____ observation deleted. If we compute the standard deviation of residual i using $s_{(i)}$ instead of s, and then compute the standardized residual for observation i using the _____ value, the resulting standardized residual is called a _____.

3. If the ith observation is an outlier, $s_{(i)}$ will be _____ than s. The absolute value of the ith studentized deleted residual therefore will be _____ the absolute value of the standardized residual.

4. Studentized deleted residuals may detect outliers that standardized residuals do not detect.

5. The t distribution can be used to determine whether the studentized deleted residuals indicate the presence of outliers.

   (a) If we delete the ith observation, the number of observations in the reduced data set is $n-1$; in this case the error sum of squares has _____ degrees of freedom.

   (b) Example For the Butler Trucking example with $n = 10$ and $p = 2$, the degrees of freedom for the error sum of squares with the ith observation deleted is $9-2-1 = 6$. At $\alpha = 0.05$ level of significance, the t distribution shows that with six degrees of freedom, _____.

   (c) If the value of the ith studentized deleted residual is _____ or _____, we can conclude that the ith observation is an outlier.

   (d) (Table 15.8) Butler Trucking example, outliers are not present in the data set.

TABLE 15.8 Studentized Deleted Residuals for Butler Trucking

Kilometers Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Standardized Residual	Studentized Deleted Residual
160	4	9.3	.78344	.75939
80	3	4.8	−.34962	−.32654
160	4	8.9	−.08334	−.07720
160	2	6.5	−1.30929	−1.39494
80	2	4.2	.38167	.35709
128	2	6.2	.65431	.62519
120	3	7.4	1.68917	2.03187
104	4	6.0	−1.77372	−2.21314
144	3	7.6	.36703	.34312
144	2	6.1	−.77639	−.75190

Influential Observations

1. (Section 14.9) we discussed how the leverage of an observation can be used to identify observations for which the value of the _____ variable may have a strong

influence on the regression results.

2. The leverage of an observation, denoted h_i, measures how far the values of the independent variables are from their mean values.

3. We use the rule of thumb _____ to identify influential observations.

4. Example Butler Trucking example ($n = 10, p = 2$)

   (a) The critical value for leverage is $3(2 + 1)/10 = 0.9$.

   (b) (Table 15.9) Because h_i does not exceed 0.9, we do not detect influential observations in the data set.

TABLE 15.9 Leverage and Cook's Distance Measures for Butler Trucking

Kilometers Traveled (x_1)	Deliveries (x_2)	Travel Time (y)	Leverage (h_i)	Cook's D (D_i)
160	4	9.3	.351704	.110994
80	3	4.8	.375863	.024536
160	4	8.9	.351704	.001256
160	2	6.5	.378451	.347923
80	2	4.2	.430220	.036663
128	2	6.2	.220557	.040381
120	3	7.4	.110009	.117562
104	4	6.0	.382657	.650029
144	3	7.6	.129098	.006656
144	2	6.1	.269737	.074217

Using Cook′ s Distance Measure to Identify

1. A problem that can arise in using leverage to identify influential observations is that an observation can be identified as having _____ and not necessarily be influential in terms of the resulting _____.

   (a) (Table 15.10) Because the leverage for the eighth observation is _____ (the critical leverage value), this observation is identified as influential.

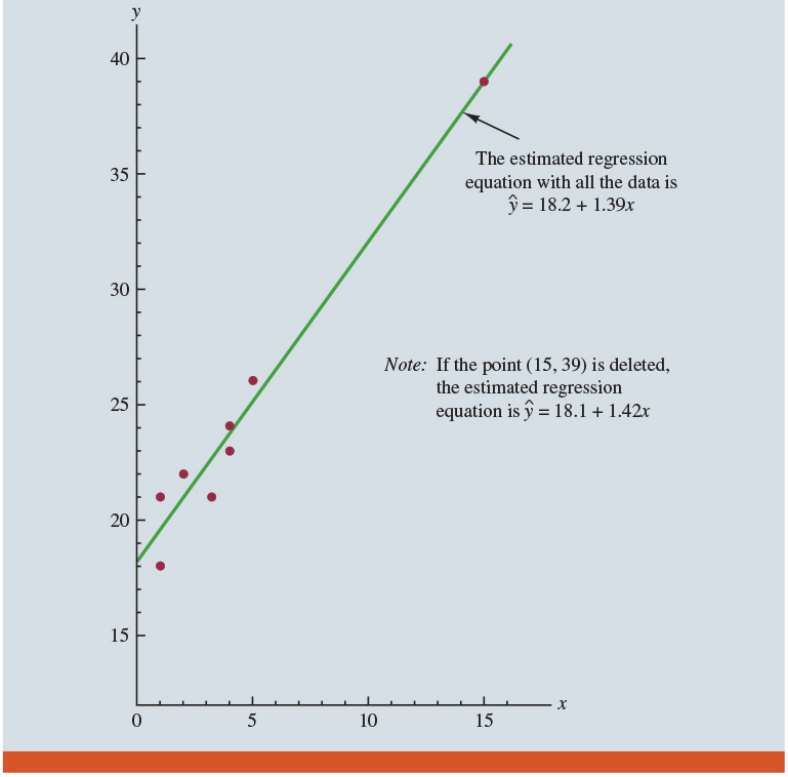
TABLE 15.10

Data Set Illustrating Potential Problem Using the Leverage Criterion

x_i	y_i	Leverage h_i
1	18	.204170
1	21	.204170
2	22	.164205
3	21	.138141
4	23	.125977
4	24	.125977
5	26	.127715
15	39	.909644

(b) (Figure 15.11) the estimated regression equation: $\hat{y} = 18.2 + 1.39x$

FIGURE 15.11 Scatter Diagram for the Data Set in Table 15.10



(c) Delete the observation $x = 15, y = 39$ from the data set and fit a new estimated regression equation to the remaining seven observations; the new estimated regression equation is $\hat{y} = 18.1 + 1.42x$

(d) We note that the y-intercept and slope of the new estimated regression equation

are very close to the values obtained using all the data.

(e) Although the leverage criterion identified the eighth observation as influential, this observation clearly had little influence on the results obtained. Thus, in some situations using only leverage to identify influential observations can lead to wrong conclusions.

2. **Cook′ s distance measure** uses both the leverage of observation i, h_i, and the residual for observation i, $(y_i - \hat{y}_i)$, to determine whether the observation is influential.

$$D_i = \frac{\qquad\qquad\qquad\qquad}{\qquad\qquad\qquad\qquad}$$

(a) The value of Cook′ s distance measure will be large and indicate an influential observation if the residual or the leverage is large.

(b) As a rule of thumb, values of _____ indicate that the ith observation is influential and should be studied further.

(c) Example (Table 15.9) Cook′ s distance measure for the Butler Trucking problem. Observation 8 with $D_i = 0.650029 < 1$, we should not be concerned about the presence of influential observations in the Butler Trucking data set.

😍 **EXERCISES 15.8**: 40, 41

15.9 Logistic Regression

1. In many regression applications, the dependent variable may only assume _____.

2. Example A bank might want to develop an estimated regression equation for predicting whether a person will be approved for a credit card. The dependent variable can be coded as _____ if the bank _____ the request for a credit card and _____ if the bank _____ the request for a credit card.

3. Using _____ we can estimate the _____ that the bank will approve the request for a credit card given a particular set of values for the chosen independent variables.

4. Example Simmons Stores. Let us consider an application of logistic regression involving a direct mail promotion being used by Simmons Stores.

   (a) Simmons owns and operates a national chain of women's apparel stores. Five thousand copies of an expensive four-color sales catalog have been printed, and each catalog includes a coupon that provides a $50 discount on purchases of $200 or more. The catalogs are expensive and Simmons would like to send them to only those customers who have a high probability of using the coupon.

   (b) Management believes that annual spending at Simmons Stores and whether a customer has a Simmons credit card are two variables that might be helpful in predicting whether a customer who receives the catalog will use the coupon.

   (c) Simmons conducted a pilot study using a random sample of 50 Simmons credit card customers and 50 other customers who do not have a Simmons credit card. Simmons sent the catalog to each of the 100 customers selected. At the end of a test period, Simmons noted whether each customer had used her or his coupon.

   (d) (Table 15.11) The amount each customer spent last year at Simmons is shown in thousands of dollars and the credit card information has been coded as 1 if the customer has a Simmons credit card and 0 if not. In the Coupon column, a 1 is recorded if the sampled customer used the coupon and 0 if not.

TABLE 15.11 Partial Sample Data for the Simmons Stores Example

Customer	Annual Spending ($1000)	Simmons Card	Coupon
1	2.291	1	0
2	3.215	1	0
3	2.135	1	0
4	3.924	0	0
5	2.528	1	0
6	2.473	0	1
7	2.384	0	0
8	7.076	0	0
9	1.182	1	1
10	3.345	0	0

(e) We might think of building a _____ model using the data in Table 15.11 to help Simmons estimate whether a catalog recipient will use the coupon. We would use Annual Spending ($1000) and Simmons Card as independent variables and Coupon as the dependent variable.

5. Because the dependent variable may only assume the values of 0 or 1, however, the _____ model is not applicable. This example shows the type of situation for which logistic regression was developed.

Logistic Regression Equation

1. In multiple regression analysis, the mean or expected value of y is referred to as the multiple regression equation.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.26)$$

2. (**Logistic Regression Equation**) In logistic regression, statistical theory as well as practice has shown that the relationship between $E(y)$ and $x_1, x_2, \cdots, x_p$ is better described by the following nonlinear equation.

$$E(y) = \_\_\_\_\_ \qquad (15.27)$$

3. If the two values of the dependent variable y are coded as 0 or 1, the value of $E(y)$ in equation (15.27) provides the _____ given a particular set of values for the independent variables $x_1, x_2, \cdots, x_p$.

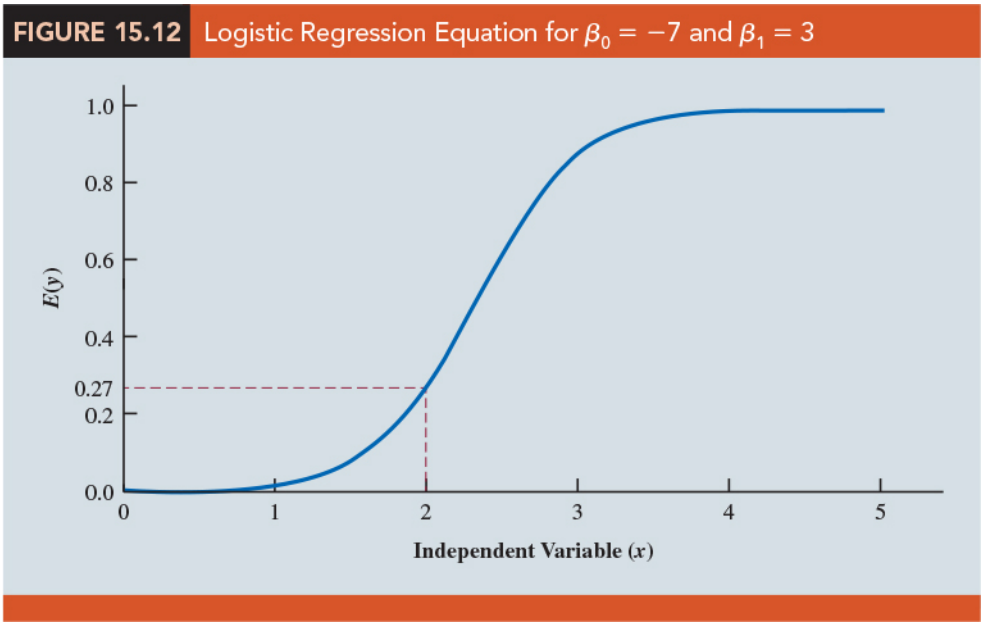
4. Because of the interpretation of $E(y)$ as a probability, the logistic regression equation is often written:

$$E(y) = \_\_\_\_\_ \qquad (15.28)$$

5. Example Suppose the model involves only one independent variable x and the values of the model parameters are $\beta_0 = -7$ and $\beta_1 = 3$. The logistic regression equation corresponding to these parameter values is

$$E(y) = \_\_\_\_\_ \qquad (15.29)$$

(a) (Figure 15.12) shows a graph of equation (15.29). Note that the graph is _____. The value of $E(y)$ ranges from _____.



(b) For example, when $x = 2$, $E(y)$ is approximately 0.27. Also note that the value of $E(y)$ gradually approaches _____ as the value of x becomes _____ and the value of $E(y)$ approaches _____ as the value of x becomes _____.

(c) For example, when $x = 2, E(y) = 0.269$. Note also that the values of $E(y)$, representing _____, increase fairly rapidly as x _____. The fact that the values of $E(y)$ range from 0 to 1 and that the curve is S-shaped makes equation (15.29) ideally suited to model the probability the dependent variable is equal to 1.

Estimating the Logistic Regression Equation

1. The _____ of the logistic regression equation makes the method of computing estimates more complex and beyond the scope of this text. We use statistical _____ to provide the estimates.

2. The estimated logistic regression equation is

$$\hat{y} = \underline{\qquad\qquad} = \underline{\qquad\qquad} \tag{15.30}$$

3. Here, $\hat{y}$ provides an _____ given a particular set of values for the independent variables.

4. Example Simmons Stores

   (a) The variables are defined:

$$y = \begin{cases} \_\_\_\_\_ & \text{if the customer did not use the coupon} \\ \_\_\_\_\_ & \text{if the customer used the coupon} \end{cases}$$

$$x_1 = \text{annual spending at Simmons Stores (\$1000s)}$$

$$x_2 = \begin{cases} \_\_\_\_\_ & \text{if the customer does not have a Simmons credit card} \\ \_\_\_\_\_ & \text{if the customer has a Simmons credit card} \end{cases}$$

   (b) Thus, we choose a logistic regression equation with two independent variables.

$$E(y) = \_\_\_\_\_ \qquad (15.31)$$

   Using the sample data (see Table 15.11), we used statistical software to compute estimates of the model parameters b_0, b_1, and b_2.

   (c) (Figure 15.13)

FIGURE 15.13 Logistic Regression Output for the Simmons Stores Example

Significance Tests

Term	Degrees of Freedom	χ^2	p-Value
Whole Model	2	13.63	.0011
Spending	1	7.56	.0060
Card	1	6.41	.0013

Parameter Estimates

Term	Estimate	Standard Error
Intercept	−2.146	.577
Spending	.342	.129
Card	1.099	.44

Odds Ratios

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

(d) We see that $\beta_0 = -2.146$, $\beta_1 = 0.342$, and $\beta_2 = 1.099$. Thus, the estimated logistic regression equation is

$$\hat{y} = \frac{e^{-2.146+0.342x_1+1.099x_2}}{1+e^{-2.146+0.342x_1+1.099x_2}} \qquad (15.32)$$

(e) An estimate the probability of using the coupon for customers who spend \$2000 annually and do not have a Simmons credit card is approximately 0.19 ($x_1 = 2$ and $x_2 = 0$):

$\hat{y} =$ _____

(f) An estimate the probability of using the coupon for customers who spent \$2000 last year and have a Simmons credit card is approximately 0.41. ($x_1 = 2$ and $x_2 = 1$):

$$\hat{y} = \frac{e^{-2.146+0.342(2)+1.099(1)}}{1+e^{-2.146+0.342(2)+1.099(1)}} = \frac{e^{-0.363}}{1+e^{-0.363}} = 0.4102$$

(g) It appears that the probability of using the coupon is _____ for customers with a Simmons credit card.

(h) Before reaching any conclusions, however, we need to assess the statistical _____.

Testing for Significance

1. Testing for significance in logistic regression is similar to testing for significance in multiple regression.

2. First we conduct a test for _____. For the Simmons Stores example, the hypotheses for the test of overall significance follow:

H_0 : _____

H_a : One or both of the parameters is not equal to zero

(a) The test for overall significance is based upon the value of a _____ statistic. If the null hypothesis is true, the sampling distribution of χ^2 follows a chi-square distribution with degrees of freedom equal to the _____ in the model.

(b) (Figure 15.13) The calculations of χ^2 is _____. The value of χ^2 and its corresponding p-value in the Whole Model row of the Significance Tests table is 13.63 and its p-value is 0.0011. Thus, at any level of significance $\alpha \geq 0.0011$, we would reject the null hypothesis and conclude that the overall model is significant.

(c) NOTE:

i. Logistic Regression: `https://online.stat.psu.edu/stat462/node/207/`

ii. Logistic regression (Wikipedia): `https://en.wikipedia.org/wiki/Logistic_regression`

3. If the χ^2 test shows an overall significance, another _____ can be used to determine whether each of the _____ independent variables is making a significant contribution to the overall model.

(a) For the independent variables x_i, the hypotheses are

_____

(b) If the null hypothesis is true, the sampling distribution of χ^2 follows a chi-square distribution with one degree of freedom.

(c) (Figure 15.13) The Spending and Card rows of the Significance Tests table of Figure 15.13 contain the values of χ^2 and their corresponding p-values test for the estimated coefficients. Suppose we use $\alpha = 0.05$ to test for the significance of the independent variables in the Simmons model.

(d) For the independent variable Spending (x_1) the χ^2 value is _____ and the corresponding p-value is _____. Thus, at the 0.05 level of significance we can reject $H_0 : \beta_1 = 0$.

(e) In a similar fashion we can also reject $H_0 : \beta_2 = 0$ because the p-value corresponding to Card's _____ is _____. Hence, at the 0.05 level of significance, both independent variables are statistically significant.

Managerial Use

1. We described how to develop the estimated logistic regression equation and how to test it for significance.

2. Example For Simmons Stores, we already computed $P(y = 1|x_1 = 2, x_2 = 1) = 0.4102$ and $P(y = 1|x_1 = 2, x_2 = 0) = 0.1881$. These probabilities indicate that for customers with annual spending of \$2000 the presence of a Simmons credit card _____ of using the coupon.

3. (Table 15.12) The estimated probabilities for values of annual spending ranging from \$1000 to \$7000 for both customers who have a Simmons credit card and customers who do not have a Simmons credit card.

TABLE 15.12 Estimated Probabilities for Simmons Stores

		Annual Spending						
		\$1000	\$2000	\$3000	\$4000	\$5000	\$6000	\$7000
Credit Card	Yes	.3307	.4102	.4948	.5796	.6599	.7320	.7936
	No	.1414	.1881	.2460	.3148	.3927	.4765	.5617

4. How can Simmons use this information to better target customers for the new promotion? Suppose Simmons wants to send the promotional catalog only to customers who have a _____ probability of using the coupon. Using the estimated probabilities in Table 15.12, Simmons promotion strategy would be:

   (a) Customers who have a Simmons credit card: Send the catalog to every customer who spent a\$2000 or more last year.

   (b) Customers who do not have a Simmons credit card: Send the catalog to every customer who spent _____ or more last year.

5. The probability of using the coupon for customers who do not have a Simmons credit card but spend \$5000 annually is _____. Thus, Simmons may want to consider revising this strategy by including those customers who _____ a credit card, as long as they spent _____ or more last year.

Interpreting the Logistic Regression Equation

1. With logistic regression, it is difficult to interpret the relation between the independent variables and the _____ directly because the logistic regression equation is _____.

2. The relationship can be interpreted indirectly using a concept called the _____(勝算比).

3. The _____(勝算) in favor of an event occurring is defined as the probability the event _____ divided by the probability the event _____. In logistic regression the event of interest is always _____.

4. Given a particular set of values for the independent variables, the odds in favor of $y = 1$ can be calculated as follows:

$$odds = \frac{\qquad\qquad}{\qquad\qquad} = \frac{\qquad\qquad}{\qquad\qquad} \tag{15.33}$$

5. The odds ratio is the odds that $y = 1$ given that one of the independent variables has been increased by _____ divided by the odds that $y = 1$ given _____ in the values for the independent variables _____.

   (a) **Odds Ratio**

$$\text{Odds Ratio} = \frac{\qquad\qquad}{\qquad\qquad} \tag{15.34}$$

   (b) For example, suppose we want to compare the odds of using the coupon for customers who spend \$2000 annually and have a Simmons credit card ($x_1 = 2$ and $x_2 = 1$) to the odds of using the coupon for customers who spend \$2000 annually and do not have a Simmons credit card ($x_1 = 2$ and $x_2 = 0$).

   (c) We are interested in interpreting the effect of a one-unit increase in the independent variable x_2. In this case

$$odd_{s1} = \frac{\qquad\qquad}{\qquad\qquad}$$

   and

$$odd_{s0} = \frac{\qquad\qquad}{\qquad\qquad}$$

   (d) Previously we showed that an estimate of the probability that $y = 1$ given $x_1 = 2$ and $x_2 = 1$ is 0.4102, and an estimate of the probability that $y = 1$ given $x_1 = 2$ and $x_2 = 0$ is 0.1881. Thus,

$$\text{estimate of } odd_{s1} = \frac{0.4102}{1 - 0.4102} = 0.6956$$

and

$$\text{estimate of } odd_{s2} = \frac{0.1881}{1 - 0.1881} = 0.2318$$

The estimated odds ratio is

$$\text{estimated odds ratio} = \frac{0.6956}{0.2318} = 3.00$$

(e) Thus, we can conclude that the _____ in favor of using the coupon for customers who spent \$2000 last year and have a Simmons credit card are _____ the estimated odds in favor of using the coupon for customers who spent \$2000 last year and do not have a Simmons credit card.

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

7. The odds ratio for each independent variable is computed while holding all the other independent variables _____. But it does not matter what constant values are used for the other independent variables. For instance, if we computed the odds ratio for the Simmons credit card variable (x_2) using \$3000, instead of \$2000, as the value for the annual spending variable (x_1), we would still obtain the _____ for the estimated odds ratio (3.00). Thus, we can conclude that the estimated odds of using the coupon for customers who have a Simmons credit card are 3 times greater than the estimated odds of using the coupon for customers who do not have a Simmons credit card.

8. (Figure 15.13) the estimated odds ratios for each of the independent variables. The estimated odds ratio for Spending (x_1) is _____ and the estimated odds ratio for Card (x_2) is _____.

9. Let us now consider the interpretation of the estimated odds ratio for the continuous independent variable x_1. The value of 1.4073 in the Odds Ratio column of the output tells us that the _____ in favor of using the coupon for customers who spent \$3000 last year is _____ the estimated odds in favor of using the coupon for customers who spent \$2000 last year.

10. A unique relationship exists between the _____ for a variable and its corresponding _____. For each independent variable in a logistic regression equation it can be shown that

---

(a) To illustrate this relationship, consider the independent variable x_1 in the Simmons example. The estimated odds ratio for x_1 is

$$\text{Estimated odds ratio } = e^{b_1} = e^{0.342} = 1.407$$

Similarly, the estimated odds ratio for x_2 is

$$\text{Estimated odds ratio } = e^{b_2} = e^{1.099} = 3.000$$

(b) 補充:

$$\begin{aligned}
\hat{p} &= \frac{e^{b_0+b_1x_1}}{1+e^{b_0+b_1x_1}}, \quad 1-\hat{p} = \frac{1}{1+e^{b_0+b_1x_1}} \\
\ln(\hat{p}) - \ln(1-\hat{p}) &= \ln(e^{b_0+b_1x_1}) - \ln(1+e^{b_0+b_1x_1}) - \ln(1) + \ln(1+e^{b_0+b_1x_1}) \\
\ln(\frac{\hat{p}}{1-\hat{p}}) &= b_0 + b_1x_1 \\
\frac{\hat{p}}{1-\hat{p}} &= e^{b_0+b_1x_1} \\
\frac{\hat{p}}{1-\hat{p}}|_{x_1=0} &= e^{b_0}, \qquad \frac{\hat{p}}{1-\hat{p}}|_{x_1=1} = e^{b_0+b_1} \\
&\Rightarrow \text{ odds ratio}|_{x=1/x=0} = e^{b_1}
\end{aligned}$$

11. The odds ratio for an independent variable represents the _____ for a _____ change in the independent variable holding all the other independent variables _____.

(a) Suppose that we want to consider the effect of a change of more than one unit, say c units. For instance, suppose in the Simmons example that we want to compare the odds of using the coupon for customers who spend \$5000 annually ($x_1 = 5$) to the odds of using the coupon for customers who spend \$2000 annually ($x_1 = 2$). In this case $c = 5-2 = 3$ and the corresponding estimated odds ratio is

---

(b) This result indicates that the estimated odds of using the coupon for customers who spend \$5000 annually is _____ greater than the estimated odds of using the coupon for customers who spend \$2000 annually.

(c) In other words, the estimated odds ratio for an increase of \$3000 in annual spending is 2.79.

(d) In general, the odds ratio enables us to compare the odds for two different events. If the value of the odds ratio is _____, the odds for both events are the same. Thus, if the independent variable we are considering (such as Simmons credit card status) has a _____ on the probability of the event occurring, the corresponding odds ratio will be _____.

12. (Figure 15.13) Most statistical software packages provide a confidence interval for the odds ratio. The Odds Ratio table in Figure 15.13 provides a 95% confidence interval for each of the odds ratios.

(a) For example, the point estimate of the odds ratio for x_1 is 1.4073 and the 95% confidence interval is _____. Because the confidence interval does not contain the value of _____, we can conclude that x_1 has a _____ relationship with the estimated odds ratio.

(b) Similarly, the 95% confidence interval for the odds ratio for x_2 is _____. Because this interval does not contain the value of 1, we can also conclude that x_2 has a significant relationship with the odds ratio.

Logit Transformation

1. It can be shown that

_____

2. This equation shows that the natural logarithm of the odds in favor of $y = 1$ is a linear function of the independent variables. This linear function is called the _____. We will use the notation _____ to denote the logit.

3. Logit

$$g(x_1, x_2, \cdots, x_p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \qquad (15.35)$$

4. Substituting $g(x_1, x_2, \cdots, x_p)$ for $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$ in equation (15.27), we can write the logistic regression equation as

$$E(y) = \underline{\qquad\qquad\qquad\qquad} \tag{15.36}$$

5. Once we estimate the parameters in the logistic regression equation, we can compute an estimate of the logit. Using $\hat{g}(x_1, x_2, \cdots, x_p)$ to denote the estimated logit, we obtain

Estimated Logit $\underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$ (15.37)

6. Thus, in terms of the estimated logit, the estimated regression equation is

$$\hat{y} = \frac{e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}}{1+e^{b_0+b_1x_1+b_2x_2+\cdots+b_px_p}} = \frac{e^{\hat{g}(x_1,x_2,\cdots,x_p)}}{1+e^{\hat{g}(x_1,x_2,\cdots,x_p)}} \tag{15.38}$$

7. For the Simmons Stores example, the estimated logit is

$\underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}$

and the estimated regression equation is

$$\hat{y} = \frac{e^{\hat{g}(x_1,x_2)}}{1+e^{\hat{g}(x_1,x_2)}} = \frac{e^{-2.146+0.342x_1+1.099x_2}}{1+e^{-2.146+0.342x_1+1.099x_2}}$$

Thus, because of the unique relationship between the estimated logit and the estimated logistic regression equation, we can compute the estimated probabilities for Simmons Stores by dividing $e^{\hat{g}(x_1,x_2)}$ by $1 + e^{\hat{g}(x_1,x_2)}$.

😍 **EXERCISES 15.9**: 44, 46, 48

😍 **SUPPLEMENTARY EXERCISES**: 51, 55.