Key Concept

This section presents a method for analyzing a linear relationship involving more than two variables.

We focus on three key elements:

- The multiple regression equation.
- The values of the adjusted $R^2$.
- The $P$-value.

Part 1: Basic Concepts of a Multiple Regression Equation

Definition

A multiple regression equation expresses a linear relationship between a response variable $y$ and two or more predictor variables $(x_1, x_2, x_3, \ldots, x_k)$.

The general form of the multiple regression equation obtained from sample data is

$$
\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k.
$$
**Notation**

\[ \hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k \]

(General form of the multiple regression equation)

- \( n \) = sample size
- \( k \) = number of predictor variables
- \( \hat{y} \) = predicted value of \( y \)
- \( x_1, x_2, x_3, \ldots, x_k \) are the predictor variables

**Notation - cont**

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \]

\( \beta_0, \beta_1, \beta_2, \ldots, \beta_k \) are the parameters for the multiple regression equation

\( b_0, b_1, b_2, \ldots, b_k \) are the sample estimates of the parameters \( \beta_0, \beta_1, \beta_2, \ldots, \beta_k \)

**Technology**

Use a statistical software package such as

- STATDISK
- Minitab
- Excel
- TI-83/84

**Example:**

Table 10-6 includes a random sample of heights of mothers, fathers, and their daughters (based on data from the National Health and Nutrition Examination). Find the multiple regression equation in which the response (y) variable is the height of a daughter and the predictor (x) variables are the height of the mother and height of the father.
Example:
The Minitab results are shown here:

```
   The regression equation is
   Height = 7.5 + 0.707 Mother + 0.164 Father

   Predictor  Coef  SE Coef   T   P
   ------    ----  ------   --  --
   Constant  7.45   10.88  0.69  0.503
   Mother    0.7072  0.1289  5.49  0.000
   Father    0.1636  0.1266  1.29  0.213

   S = 1.93990    R-Sq = 67.5%   R-Sq(adj) = 63.7%
```

Analysis of Variance

```
   Source     DF    SS     MSE    F    P
   Regression  2  132.997  66.499  17.67  0.000
   Residual Error 17  63.975  3.763
   Total        19 196.972
```

Example:
From the display, we see that the multiple regression equation is

\[ \hat{y} = 7.5 + 0.707x_1 + 0.164x_2 \]

where \( \hat{y} \) is the predicted height of a daughter, \( x_1 \) is the height of the mother, and \( x_2 \) is the height of the father.

Definition

- The multiple coefficient of determination \( R^2 \) is a measure of how well the multiple regression equation fits the sample data.

- The adjusted coefficient of determination is the multiple coefficient of determination \( R^2 \) modified to account for the number of variables and the sample size.

Adjusted Coefficient of Determination

\[
\text{Adjusted } R^2 = 1 - \frac{(n-1)}{[n-(k+1)]} (1-R^2)
\]

where \( n \) = sample size
\( k \) = number of predictor (x) variables

**Formula 10-7**
**P-Value**

The *P*-value is a measure of the overall significance of the multiple regression equation. Like the adjusted $R^2$, this *P*-value is a good measure of how well the equation fits the sample data.

The displayed Minitab *P*-value of 0.000 (rounded to three decimal places) is small, indicating that the multiple regression equation has good overall significance and is usable for predictions. That is, it makes sense to predict heights of daughters based on heights of mothers and fathers. The value of 0.000 results from a test of the null hypothesis that $\beta_1 = \beta_2 = 0$. Rejection of $\beta_1 = \beta_2 = 0$ implies that at least one of $\beta_1$ and $\beta_2$ is not 0, indicating that this regression equation is effective in predicting heights of daughters.

**Finding the Best Multiple Regression Equation**

1. Use common sense and practical considerations to include or exclude variables.

2. Consider the *P*-value. Select an equation having overall significance, as determined by the *P*-value found in the computer display.

3. Consider equations with high values of adjusted $R^2$ and try to include only a few variables.
   - If an additional predictor variable is included, the value of adjusted $R^2$ does not increase by a substantial amount.
   - For a given number of predictor ($x$) variables, select the equation with the largest value of adjusted $R^2$.
   - In weeding out predictor ($x$) variables that don’t have much of an effect on the response ($y$) variable, it might be helpful to find the linear correlation coefficient $r$ for each of the paired variables being considered.
Part 2: Dummy Variables and Logistic Equations

Dummy Variable

Many applications involve a dichotomous variable which has only two possible discrete values (such as male/female, dead/alive, etc.). A common procedure is to represent the two possible discrete values by 0 and 1, where 0 represents “failure” and 1 represents success.

A dichotomous variable with the two values 0 and 1 is called a dummy variable.

Logistic Regression

We can use the methods of this section if the dummy variable is the predictor variable.

If the dummy variable is the response variable we need to use a method known as logistic regression.

As the name implies logistic regression involves the use of natural logarithms. This textbook does not include detailed procedures for using logistic regression.

Recap

In this section we have discussed:

- The multiple regression equation.
- Adjusted $R^2.$
- Finding the best multiple regression equation.
- Dummy variables and logistic regression.
• Use common sense and practical considerations
• Consider the P-value.
• Consider equations with high values of adjusted $R^2$ and try to include only a few variables.
• For a given number of predictor ($x$) variables, select the equation with the largest value of adjusted $R^2$.
• In weeding out predictor ($x$) variables that don’t have much of an effect on the response ($y$) variable, it might be helpful to find the linear correlation coefficient $r$ for each of the paired variables being considered.

Home Selling Prices: Finding the Best Multiple Regression Equation. In Exercises 9–12, refer to the accompanying table, which was obtained using data from homes sold (from Data Set 23 in Appendix B). The response ($y$) variable is selling price (in dollars). The predictor ($x$) variables are LP (list price in dollars), LA (living area of the home in square feet), and LOT (lot size in acres).

572/10. If exactly two predictor ($x$) variables are to be used to predict the selling price of a home, which two variables should be chosen? Why?

572/12. A home is for sale with a list price of $400,000, it has a living area of 3000 square feet, and it is on a 2 acre lot. What is the best predicted value of the selling price? Is that predicted selling price likely to be a good estimate? Is that predicted value likely to be very accurate?

Home Selling Prices: Finding the Best Multiple Regression Equation. In Exercises 9–12, refer to the accompanying table, which was obtained using data from homes sold (from Data Set 23 in Appendix B). The response ($y$) variable is selling price (in dollars). The predictor ($x$) variables are LP (list price in dollars), LA (living area of the home in square feet), and LOT (lot size in acres).

572/11. A home is for sale with a list price of $400,000, it has a living area of 3000 square feet, and it is on a 2 acre lot. What is the best predicted value of the selling price? Is that predicted selling price likely to be a good estimate? Is that predicted value likely to be very accurate?

572/14. Predicting Movie Gross Amount Refer to Data Set 9 in Appendix B and find the best regression equation with movie gross amount (in millions of dollars) as the response ($y$) variable. Ignore the MPAA ratings. Why is this equation best? Is this “best” equation good for predicting the amount of money that a movie will gross? Does the combination of predictor variables make sense?
Predicting Movie Gross Amount
Refer to Data Set 9 in Appendix B and find the best regression equation with movie gross amount (in millions of dollars) as the response (y) variable. Ignore the MPAA ratings. Why is this equation best? Is this “best” equation good for predicting the amount of money that a movie will gross? Does the combination of predictor variables make sense?