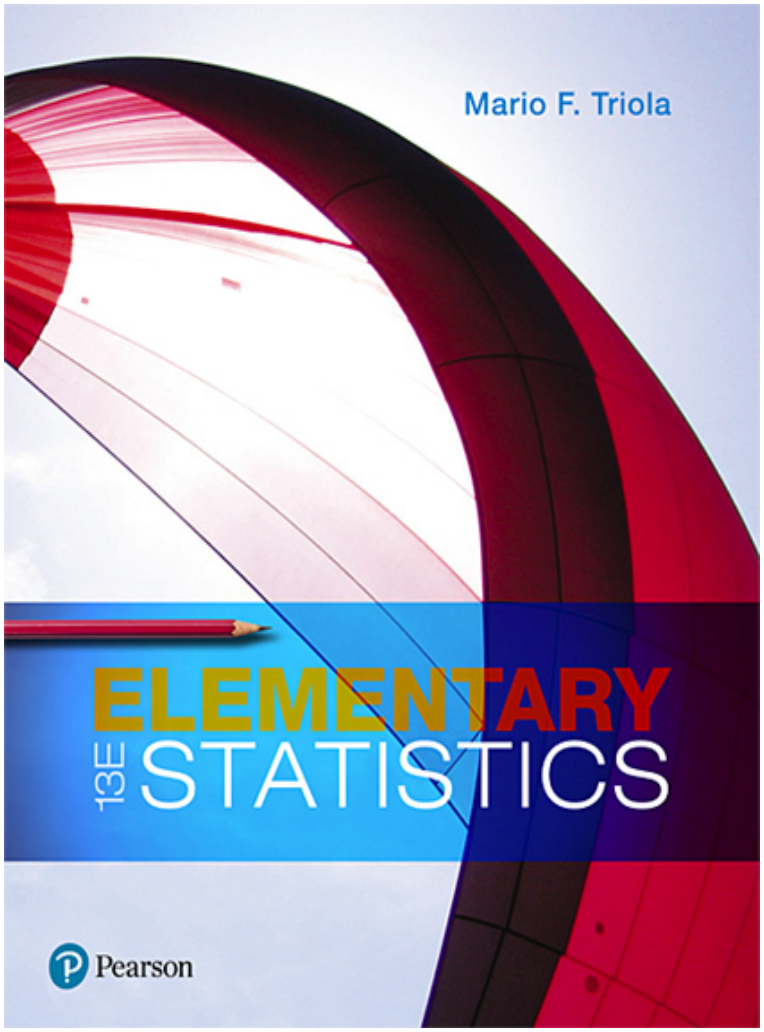
Elementary Statistics

Thirteenth Edition



Chapter 10

Correlation and Regression

Correlation and Regression

10-1 Correlation

10-2 Regression

10-3 Prediction Intervals and Variation

10-4 Multiple Regression

10-5 Nonlinear Regression

Key Concept

This section presents methods for analyzing a linear relationship with **more than two** variables. We focus on these two key elements: (1) finding the multiple regression equation, and (2) using the value of adjusted R^2 and the P-value as measures of how well the multiple regression equation fits the sample data. This section emphasizes the use and interpretation of results from technology.

Multiple Regression Equation

- Multiple Regression Equation
  - A **multiple regression equation** expresses a linear relationship between a response variable *y* and two or more predictor variables ($x_1, x_2, \ldots, x_k$). The general form of a multiple regression equation obtained from sample data is

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

Finding a Multiple Regression Equation: Objective

Use sample matched data from three or more variables to find a multiple regression equation that is useful for predicting values of the response variable *y*.

Finding a Multiple Regression Equation: Notation

$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \ldots + b_k\, x_k$ (multiple regression equation found from **sample** data)

$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_k\, x_k$ (multiple regression equation for the **population** of data)

$\hat{y}$ = predicted value of y

k = number of **predictor** variables (also called **independent** variables or x variables)

n = sample size

Finding a Multiple Regression Equation: Requirements

For any specific set of *x* values, the regression equation is associated with a random error often denoted by ε. We assume that such errors are normally distributed with a mean of 0 and a standard deviation of σ and that the random errors are independent.

Finding a Multiple Regression Equation: Procedure for Finding a Multiple Regression Equation

Manual calculations are not practical, so technology must be used.

Example: Predicting Weight (1 of 3)

Data Set 1 “Body Data” in Appendix B includes heights (cm), waist circumferences (cm), and weights (kg) from a sample of 153 males. Find the multiple regression equation in which the response variable (*y*) is the weight of a male and the predictor variables are height (x_1) and waist circumference (x_2).

Example: Predicting Weight (2 of 3)

Solution

Using Statdisk with the sample data in Data Set 1, we obtain the results shown in the display on the next slide. The coefficients b_0, b_1, and b_2 are used in the multiple regression equation:

$$\hat{y} = -149 + 0.769x_1 + 1.01x_2$$

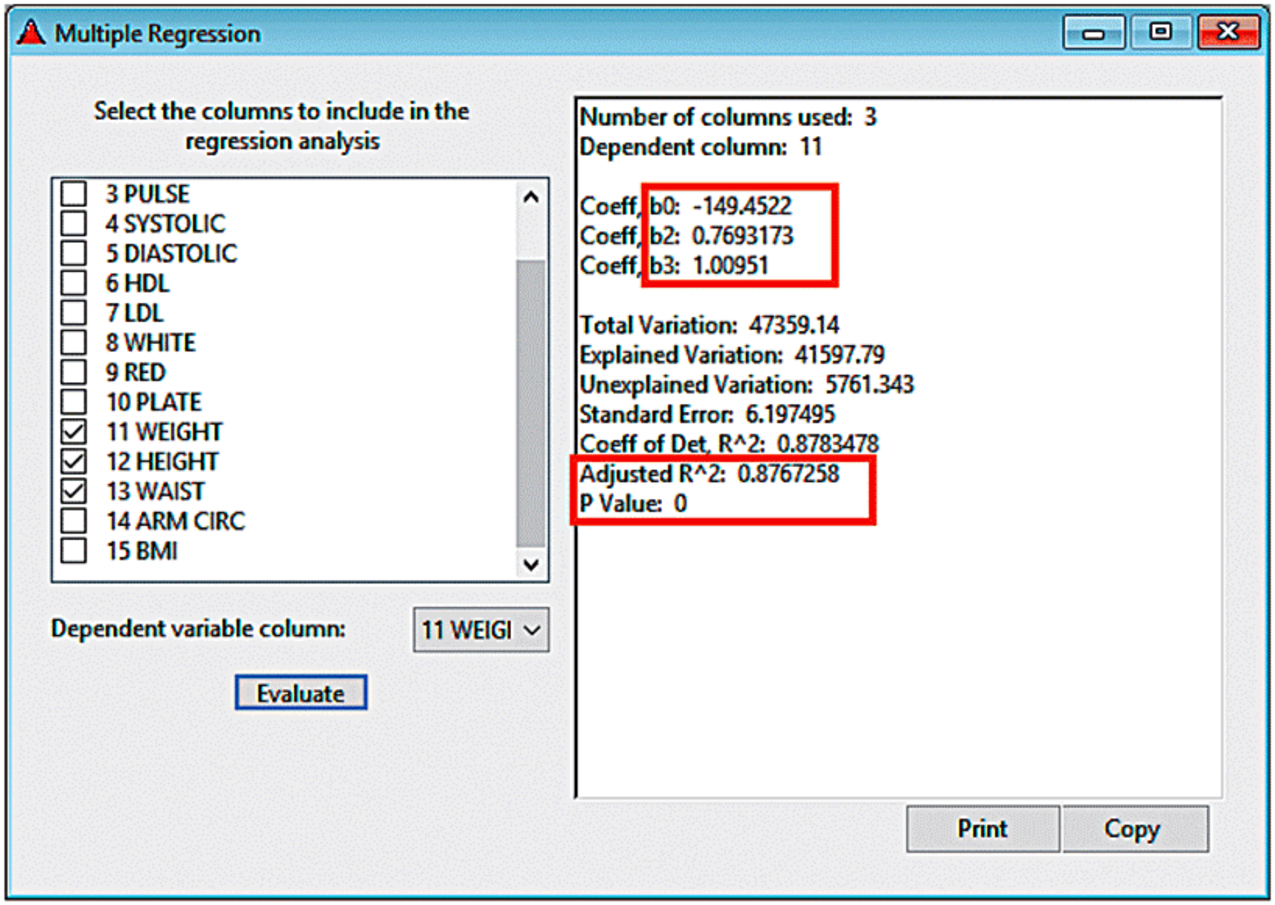
or Weight = −149 + 0.769 Height + 1.01 Waist

The obvious advantage of the second format above is that it is easier to keep track of the roles that the variables play.

Example: Predicting Weight (3 of 3)

Solution

Statdisk



Select the columns to include in the regression analysis

- ☐ 3 PULSE
- ☐ 4 SYSTOLIC
- ☐ 5 DIASTOLIC
- ☐ 6 HDL
- ☐ 7 LDL
- ☐ 8 WHITE
- ☐ 9 RED
- ☐ 10 PLATE
- ☑ 11 WEIGHT
- ☑ 12 HEIGHT
- ☑ 13 WAIST
- ☐ 14 ARM CIRC
- ☐ 15 BMI

Dependent variable column: 11 WEIGI

```
Number of columns used:  3
Dependent column:  11

Coeff, b0:  -149.4522
Coeff, b2:  0.7693173
Coeff, b3:  1.00951

Total Variation:  47359.14
Explained Variation:  41597.79
Unexplained Variation:  5761.343
Standard Error:  6.197495
Coeff of Det, R^2:  0.8783478
Adjusted R^2:  0.8767258
P Value:  0
```

Adjusted Coefficient of Determination

- Adjusted Coefficient of Determination
  - 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. It is calculated by using:

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

where

n = sample size

k = number of predictor (x) variables

Guidelines for Finding the Best Multiple Regression Equation (1 of 2)

1. Use common sense and practical considerations to include or exclude variables.
2. Consider the *P*-value.
3. Consider equations with high values of adjusted R^2, and try to include only a few variables.

Guidelines for Finding the Best Multiple Regression Equation (2 of 2)

Instead of including almost every available variable, try to include relatively few predictor (*x*) variables. Use these guidelines:

- Select an equation having a value of adjusted R^2 with this property: If an additional predictor variable is included, the value of adjusted R^2 does not increase very much.
- For a particular number of predictor (*x*) variables, select the equation with the largest value of adjusted R^2.
- In excluding 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 pair of variables being considered. If two predictor values have a very high linear correlation coefficient (called **multicollinearity**), there is no need to include them both, and we should exclude the variable with the lower value of adjusted R^2.

Example: Predicting Height from Footprint Evidence (1 of 8)

Data Set 2 “Foot and Height” in Appendix B includes the age, foot length, shoe print length, shoe size, and height for each of 40 different subjects. Using those sample data, find the regression equation that is best for predicting height. Is the “best” regression equation a **good** equation for predicting height?

Example: Predicting Height from Footprint Evidence (2 of 8)

Solution

Using the response variable of height and possible predictor variables of age, foot length, shoe print length, and shoe size, there are 15 different possible combinations of predictor variables. The table on the next slide includes key results from five of those combinations.

Example: Predicting Height from Footprint Evidence (3 of 8)

Solution

Predictor Variables	**Adjusted R^2**	**P-Value**	
Age	0.1772	0.004	← **Not best:** Adjusted R^2 is far less than 0.7014 for foot Length.
Foot Length	**0.7014**	**0.000**	← **Best:** High adjusted R^2 and lowest *P*-value.
Shoe Print Length	0.6520	0.000	← **Not best:** Adjusted R^2 is less than 0.7014 for foot Length.
Foot Length/Shoe Print Length	0.7484	0.000	← **Not best:** The adjusted R^2 value is not very much higher than 0.7014 for the single variable of Foot Length.
Age/Foot Length/Shoe Print Length/Shoe Size	0.7585	0.000	← **Not best:** There are other cases using fewer variables with adjusted R^2 that are not too much smaller.

Example: Predicting Height from Footprint Evidence (4 of 8)

Solution

Blind and thoughtless application of regression methods would suggest that the best regression equation uses all four of the predictor variables, because that combination yields the highest adjusted R^2 value of 0.7585. However, given the objective of using evidence to estimate the height of a suspect, we use **critical thinking** as follows.

Example: Predicting Height from Footprint Evidence (5 of 8)

Solution

1. Delete the variable of age, because criminals rarely leave evidence identifying their ages.
2. Delete the variable of shoe size, because it is really a rounded form of foot length.
3. For the remaining variables of foot length and shoe print length, use only foot length because its adjusted R^2 value of 0.7014 is greater than 0.6520 for shoe print length, and it is not very much less than the adjusted R^2 value of 0.7484 for both foot length and shoe print length. In this case, it is better to use one predictor variable instead of two.

Example: Predicting Height from Footprint Evidence (6 of 8)

Solution

4. Although it appears that the use of the single variable of foot length is best, we also note that criminals usually wear shoes, so shoe print lengths are more likely to be found than foot lengths.

Example: Predicting Height from Footprint Evidence (7 of 8)

Interpretation

Blind use of regression methods suggests that when estimating the height of a subject, we should use all of the available data by including all four predictor variables of age, foot length, shoe print length, and shoe size, but other practical considerations suggest that it is best to use the single predictor variable of foot length. So the best regression equation appears to be this: Height = 64.1 + 4.29 (Foot Length).

Example: Predicting Height from Footprint Evidence (8 of 8)

Interpretation

However, given that criminals usually wear shoes, it is best to use the single predictor variable of shoe print length, so the best practical regression equation appears to be this: Height = 80.9 + 3.22 (Shoe Print Length). The *P*-value of 0.000 suggests that the regression equation yields a good model for estimating height.

Dummy Variable (1 of 2)

- Dummy Variable
  - A **dummy variable** is a variable having only the values of 0 and 1 that are used to represent the two different categories of a qualitative variable.

Dummy Variable (2 of 2)

A dummy variable is sometimes called a **dichotomous variable.** The word “dummy” is used because the variable does not actually have any quantitative value, but we use it as a substitute to represent the different categories of the qualitative variable.

Example: Using a Dummy Variable as a Predictor Variable (1 of 6)

The table on the next slide is adapted from Data Set 5 “Family Heights” in Appendix B and it is in a more convenient format for this example. Use the dummy variable of sex (coded as 0 = female, 1 = male). Given that a father is 69 in. tall and a mother is 63 in. tall, find the multiple regression equation and use it to predict the height of (a) a daughter and (b) a son.

Example: Using a Dummy Variable as a Predictor Variable (2 of 6)

Height of Father	Height of Mother	Height of Child	Sex of Child (1 = Male)
66.5	62.5	70.0	1
70.0	64.0	68.0	1
67.0	65.0	69.7	1
68.7	70.5	71.0	1
69.5	66.0	71.0	1
70.0	65.0	73.0	1
69.0	66.0	70.0	1
68.5	67.0	73.0	1
65.5	60.0	68.0	1

Example: Using a Dummy Variable as a Predictor Variable (3 of 6)

Height of Father	Height of Mother	Height of Child	Sex of Child (1 = Male)
69.5	66.5	70.5	1
70.5	63.0	64.5	0
71.0	65.0	62.0	0
70.5	62.0	60.0	0
66.0	66.0	67.0	0
68.0	61.0	63.5	0
68.0	63.0	63.0	0
71.0	62.0	64.5	0
65.5	63.0	63.5	0
64.0	60.0	60.0	0
71.0	63.0	63.5	0

Example: Using a Dummy Variable as a Predictor Variable (4 of 6)

Solution

Using the methods of multiple regression from Part 1 of this section and computer software, we get this regression equation:
Height of child = 36.5 − 0.0336 (Height of father) + 0.461 (Height of mother) + 6.14 (Sex) where the value of the dummy variable of sex is either 0 for a daughter or 1 for a son.

Example: Using a Dummy Variable as a Predictor Variable (5 of 6)

Solution

a. To find the predicted height of a **daughter,** we substitute 0 for the sex variable, and we also substitute 69 in. for the father’s height and 63 in. for the mother’s height. The result is a predicted height of 63.2 in. for a daughter.

b. To find the predicted height of a **son**, we substitute 1 for the sex variable, and we also substitute 69 in. for the father’s height and 63 in. for the mother’s height. The result is a predicted height of 69.4 in. for a son.

Example: Using a Dummy Variable as a Predictor Variable (6 of 6)

Solution

The coefficient of 6.14 in the regression equation shows that when given the height of a father and the height of a mother, a son will have a predicted height that is 6.14 in. more than the height of a daughter.