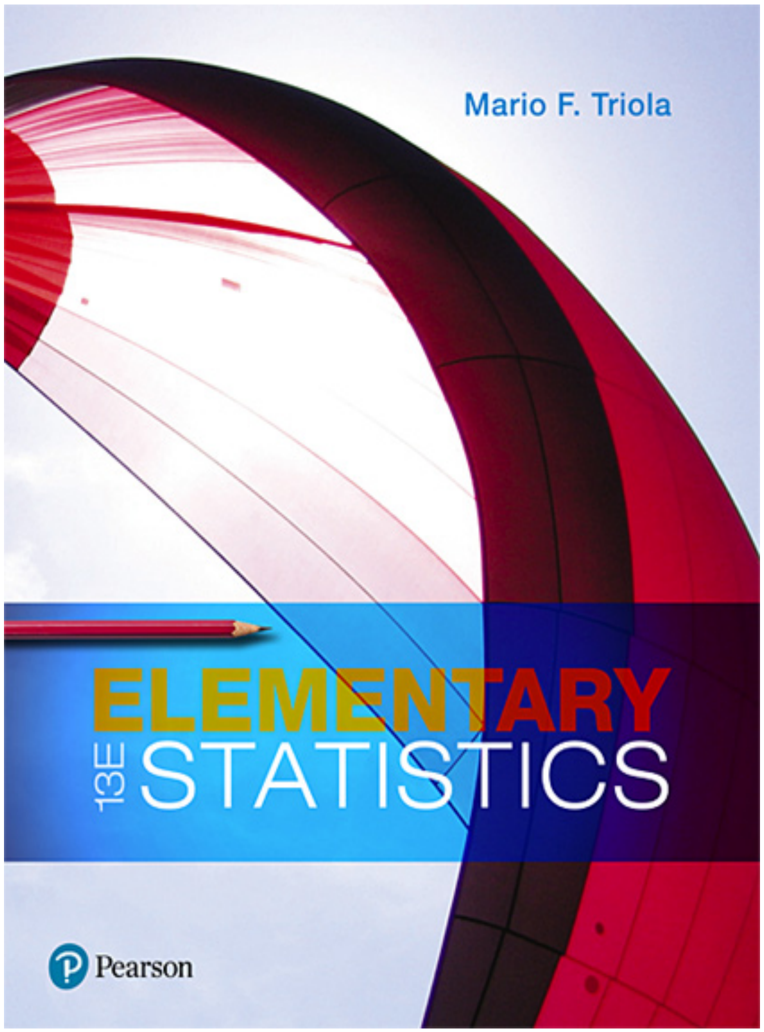
Elementary Statistics

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 Objective

In this section we introduce the **prediction interval**, which is an interval estimate of a predicted value of *y*.

Prediction Interval

- **Prediction Interval**
  - A **prediction interval** is a range of values used to estimate a **variable** (such as a predicted value of *y* in a regression equation).

Confidence Interval

- **Confidence Interval**
  - A **confidence interval** is a range of values used to estimate a population **parameter** (such as p or μ or σ).

Prediction Intervals: Objective

Find a prediction interval, which is an interval estimate of a predicted value of *y*.

Prediction Intervals: Requirement

For each fixed value of *x*, the corresponding sample values of *y* are normally distributed about the regression line, and those normal distributions have the same variance.

Prediction Intervals: Formulas for Creating a Prediction Interval (1 of 2)

Given a fixed and known value x_0, the prediction interval for an individual y value is

$$\hat{y} - E < y < \hat{y} + E$$

where the margin of error is

$$E = t_{\frac{\alpha}{2}} s_e \sqrt{1 + \frac{1}{n} + \frac{n\left(x_0 - \bar{x}\right)^2}{n\left(\sum x^2\right) - \left(\sum x\right)^2}}$$

and x_0 is a given value of x, $t_{\frac{\alpha}{2}}$ has $n - 2$ degrees of freedom, and s_e is the **standard error of estimate** found from the next formulas.

Prediction Intervals: Formulas for Creating a Prediction Interval (2 of 2)

FORMULA 10-5 $$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

Below is an equivalent form that is good for manual calculations or writing computer programs.

FORMULA 10-6 $$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

Example: Chocolate and Nobel Laureates-Finding a Prediction Interval (1 of 5)

For the paired chocolate/Nobel data previously used, we found that there is sufficient evidence to support the claim of a linear correlation between those two variables, and the regression equation is $\hat{y} = -3.37 + 2.49x$.

a. If a country has a chocolate consumption amount given by $x = 10$ kg per capita, find the best predicted value of the Nobel Laureate rate.

b. Use a chocolate consumption amount of $x = 10$ kg per capita to construct a 95% prediction interval for the Nobel Laureate rate.

Example: Chocolate and Nobel Laureates-Finding a Prediction Interval (2 of 5)

Solution

a. Substitute $x = 10$ into the regression equation $\hat{y} = -3.37 + 2.49x$ to get a predicted value of $\hat{y} = 21.5$ Nobel Laureates per 10 million people.

b. The accompanying StatCrunch and Minitab displayed on the next slide provide the 95% prediction interval, which is $7.8 < y < 35.3$ when rounded.

Example: Chocolate and Nobel Laureates-Finding a Prediction Interval (3 of 5)

Solution

StatCrunch

Predicted values:

X value	Pred. Y	s.e.(Pred. y)	95% C.I. for mean	95% P.I. for new
10	21.56467	2.1502909	(17.092895, 26.036445)	(7.7944348, 35.334905)

Minitab

Prediction for Nobel

```
Regression Equation
Nobel = -3.37 + 2.493 Chocolate

Variable    Setting
Chocolate        10

    Fit   SE Fit          95% CI                  95% PI
21.5647  2.15029  (17.0929, 26.0364)   (7.79443, 35.3349)
```

Example: Chocolate and Nobel Laureates-Finding a Prediction Interval (4 of 5)

Solution

The same 95% prediction interval could be manually calculated using these components:

$x_0 = 10$ (given)

$s_e = 6.262665$ (provided by many technologies)

$\hat{y} = 21.5$ (predicted value of y)

$t_{\frac{\alpha}{2}} = 2.080$ (from Table A-3 with df = 21 and area of 0.05 in two tails), $n = 23$,

$\bar{x} = 5.804348$, $\sum x = 133.5$, $\sum x^2 = 1011.45$

Example: Chocolate and Nobel Laureates- Finding a Prediction Interval (5 of 5)

Interpretation

The 95% prediction interval is $7.8 < y < 35.3$. This means that if we select some country with a chocolate consumption rate of 10 kg per capita ($x = 10$), we have 95% confidence that the limits of 7.8 and 35.3 contain the Nobel Laureate rate. That is a wide range of values. The prediction interval would be much narrower and our estimated Nobel rate would be much better if we were using a much larger set of sample data instead of using only the 23 pairs of values.

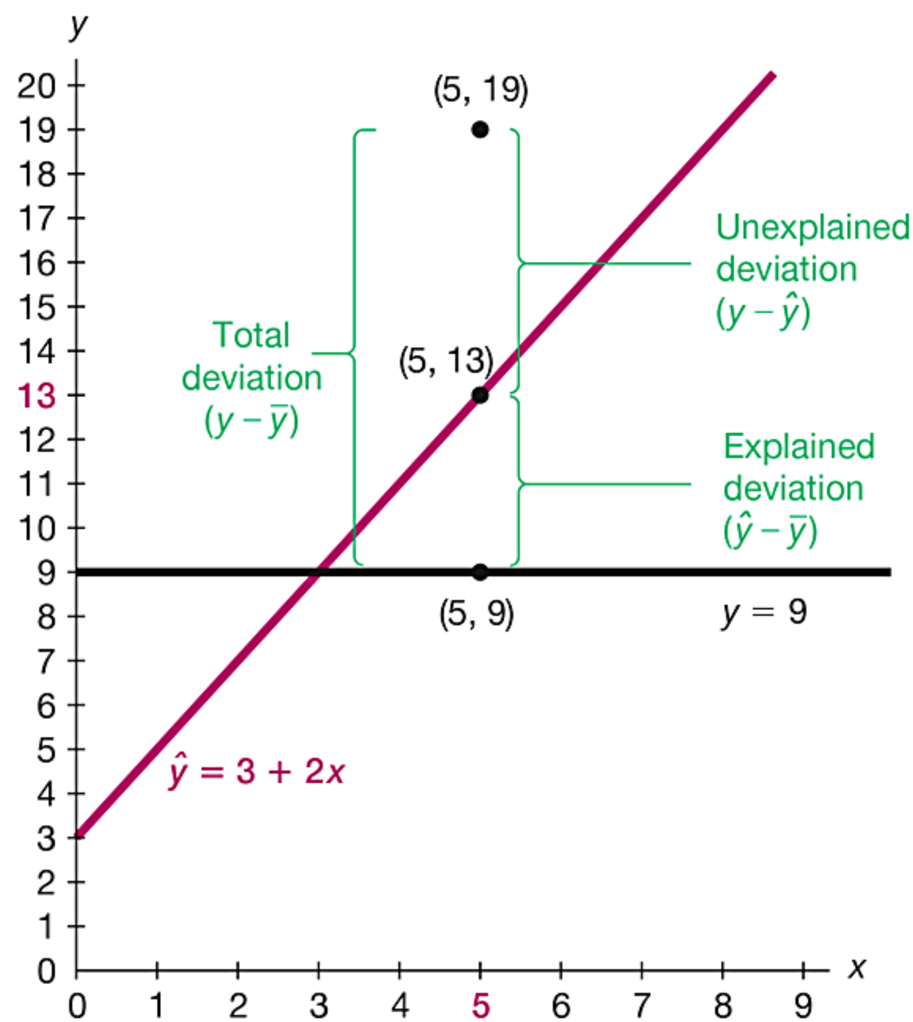
Explained and Unexplained Variation (1 of 3)

Assume that we have a sample of paired data having the following properties on the next slide:

- There is sufficient evidence to support the claim of a linear correlation between *x* and *y*.
- The equation of the regression line is $\hat{y} = 3 + 2x$.
- The mean of the *y* values is given by $\bar{y} = 9$.
- One of the pairs of sample data is $x = 5$ and $y = 19$.
- The point (5, 13) is one of the points on the regression line, because substituting $x = 5$ into the regression equation of $\hat{y} = 3 + 2x$ yields $\hat{y} = 13$.

Explained and Unexplained Variation (2 of 3)

Total, Explained, and Unexplained Deviation



Explained and Unexplained Variation (3 of 3)

Total deviation (from $\bar{y} = 9$) of the point (5, 19)

$$= y - \bar{y} = 19 - 9 = 10$$

Explained deviation (from $\bar{y} = 9$) of the point (5, 19)

$$= \hat{y} - \bar{y} = 13 - 9 = 4$$

Unexplained deviation (from $\bar{y} = 9$) of the point (5, 19)

$$= y - \hat{y} = 19 - 13 = 6$$

Total Deviation

- **Total Deviation**

  The **total deviation** of (x, y) is the vertical distance $y - \bar{y}$, which is the distance between the point (x, y) and the horizontal line passing through the sample mean $\bar{y}$.

Explained Deviation

- **Explained Deviation**

  The **explained deviation** is the vertical distance $\hat{y} - \bar{y}$, which is the distance between the predicted y value and the horizontal line passing through the sample mean $\bar{y}$.

Unexplained Deviation

- **Unexplained Deviation**

  The **unexplained deviation,** also called a **residual,** is the vertical distance $y - \hat{y}$, which is the vertical distance between the point (x, y) and the regression line.

Total Variation

(total variation) = (explained variation) + (unexplained variation)

$$\sum(y-\bar{y})^2 = \sum(\hat{y}-\bar{y})^2 + \sum(y-\hat{y})^2$$

Coefficient of Determination

- **Coefficient of Determination**

  The **coefficient of determination** is the proportion of the variation in *y* that is explained by the regression line. It is computed as

$$r^2 = \frac{\text{explained variation}}{\text{total variation}}$$

  The value of r^2 is the proportion of the variation in *y* that is explained by the linear relationship between *x* and *y*.

Example: Chocolate/Nobel Data- Finding a Coefficient of Determination (1 of 3)

If we use the 23 pairs of chocolate/Nobel data, we find that the linear correlation coefficient is $r = 0.801$. Find the coefficient of determination. Also, find the percentage of the total variation in y (Nobel rate) that can be explained by the linear correlation between chocolate consumption and Nobel rate.

Example: Chocolate/Nobel Data- Finding a Coefficient of Determination (2 of 3)

Solution

With $r = 0.801$ the coefficient of determination is $r^2 = 0.642$.

Example: Chocolate/Nobel Data- Finding a Coefficient of Determination (3 of 3)

Interpretation

Because r^2 is the proportion of total variation that can be explained, we conclude that 64.2% of the total variation in the Nobel rate can be explained by chocolate consumption, and the other 35.8% cannot be explained by chocolate consumption. The other 35.8% might be explained by some other factors and/or random variation. But common sense suggests that it is somewhat silly to seriously think that a country's rate of Nobel Laureates is affected by the amount of chocolate consumed.