Prediction Intervals and Variation

Colby Community College

Definition

A **prediction interval** is a range of values used to estimate a variable.

Definition

A **prediction interval** is a range of values used to estimate a variable.

Definition

A **confidence interval** is a range of values used to estimate a population parameter.

Definition

A **prediction interval** is a range of values used to estimate a variable.

Definition

A **confidence interval** is a range of values used to estimate a population parameter.

Formula

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

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}} \text{ and } s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

and $t_{\alpha/2}$ has $n - 2$ degrees of freedom.

Example 1

For the paired data in Data Set 16 comparing chocolate consumption and Nobel laureate rate has linear correlation and the regression equation is $\hat{y} = -3.37 + 2.49x$.

Example 1

For the paired data in Data Set 16 comparing chocolate consumption and Nobel laureate rate has linear correlation and the regression equation is $\hat{y} = -3.37 + 2.49x$.

Using the regression equation to predict the Nobel laureate rate for a country that consumes 10kg per capita is 21.5 Nobel laureates per 10 million people.

Example 1

For the paired data in Data Set 16 comparing chocolate consumption and Nobel laureate rate has linear correlation and the regression equation is $\hat{y} = -3.37 + 2.49x$.

Using the regression equation to predict the Nobel laureate rate for a country that consumes 10kg per capita is 21.5 Nobel laureates per 10 million people.

Technology gives a 95% prediction interval of $7.8 < y < 35.3$.

Example 1

For the paired data in Data Set 16 comparing chocolate consumption and Nobel laureate rate has linear correlation and the regression equation is $\hat{y} = -3.37 + 2.49x$.

Using the regression equation to predict the Nobel laureate rate for a country that consumes 10kg per capita is 21.5 Nobel laureates per 10 million people.

Technology gives a 95% prediction interval of $7.8 < y < 35.3$.

This means that if we select some country with a chocolate consumption rate of 10kg per capita, we have a 95% confidence that the limits of 7.8 and 35.3 contain the Nobel laureate rate.

Example 1

For the paired data in Data Set 16 comparing chocolate consumption and Nobel laureate rate has linear correlation and the regression equation is $\hat{y} = -3.37 + 2.49x$.

Using the regression equation to predict the Nobel laureate rate for a country that consumes 10kg per capita is 21.5 Nobel laureates per 10 million people.

Technology gives a 95% prediction interval of $7.8 < y < 35.3$.

This means that if we select some country with a chocolate consumption rate of 10kg per capita, we have a 95% confidence that the limits of 7.8 and 35.3 contain the Nobel laureate rate.

Note

We could narrow down the interval by using a much larger set of data.

Note

For the following definitions, we assume that we have a collection of paired data containing the sample point (x, y), that $\hat{y}$ is the predicted value of y (obtained by using the regression equation), and that the mean of the sample y values is $\bar{y}$.

Note

For the following definitions, we assume that we have a collection of paired data containing the sample point (x, y), that $\hat{y}$ is the predicted value of y (obtained by using the regression equation), and that the mean of the sample y values is $\bar{y}$.

Definition

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 $\bar{y}$.

Note

For the following definitions, we assume that we have a collection of paired data containing the sample point (x, y), that $\hat{y}$ is the predicted value of y (obtained by using the regression equation), and that the mean of the sample y values is $\bar{y}$.

Definition

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 $\bar{y}$.

Definition

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 $\bar{y}$.

Note

For the following definitions, we assume that we have a collection of paired data containing the sample point (x, y), that $\hat{y}$ is the predicted value of y (obtained by using the regression equation), and that the mean of the sample y values is $\bar{y}$.

Definition

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 $\bar{y}$.

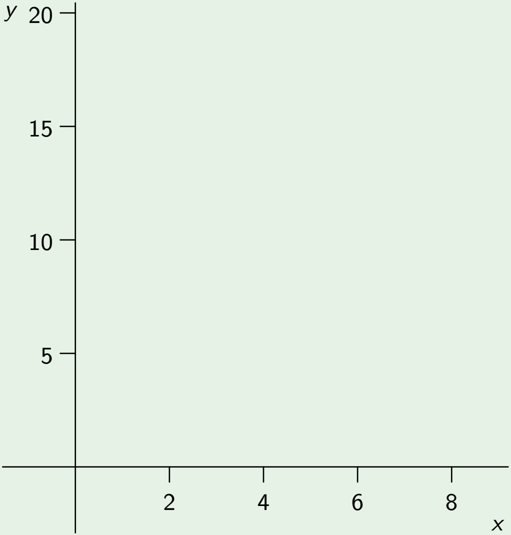
Definition

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 $\bar{y}$.

Definition

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

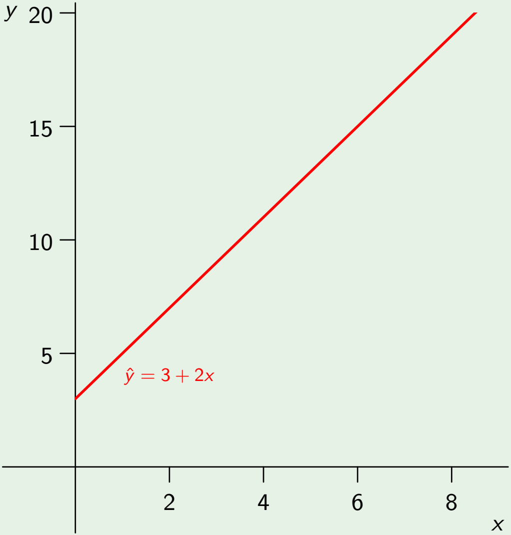
Example 2



We assume the following:

- There is sufficient evidence to support the claim of a linear correlation between x and y.

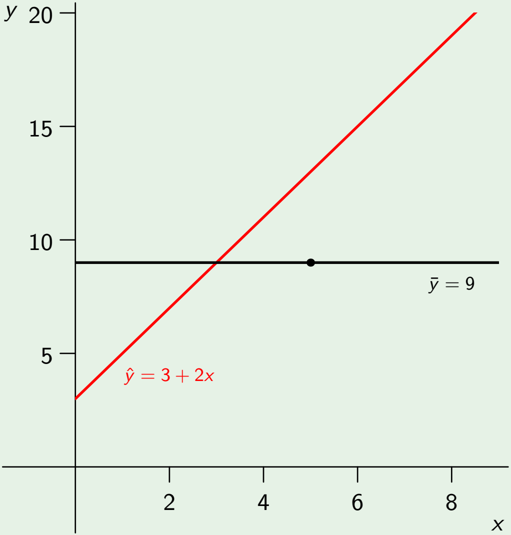
Example 2



We assume the following:

- There is sufficient evidence to support the claim of a linear correlation between x and y.
- The equation of the regression is $\hat{y} = 3 + 2x$.

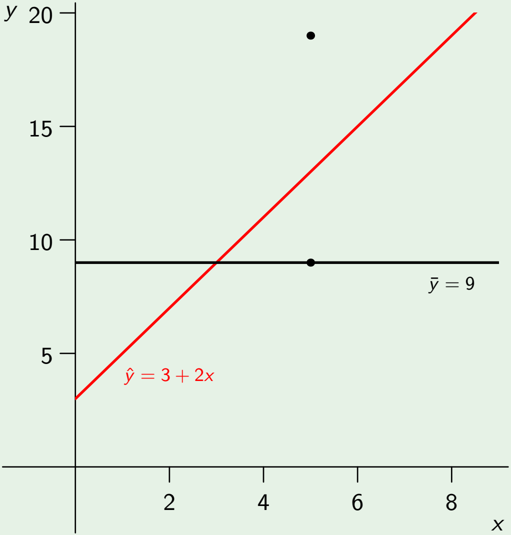
Example 2



We assume the following:

- There is sufficient evidence to support the claim of a linear correlation between x and y.
- The equation of the regression is $\hat{y} = 3 + 2x$.
- The mean of the y values is given by $\bar{y} = 9$.

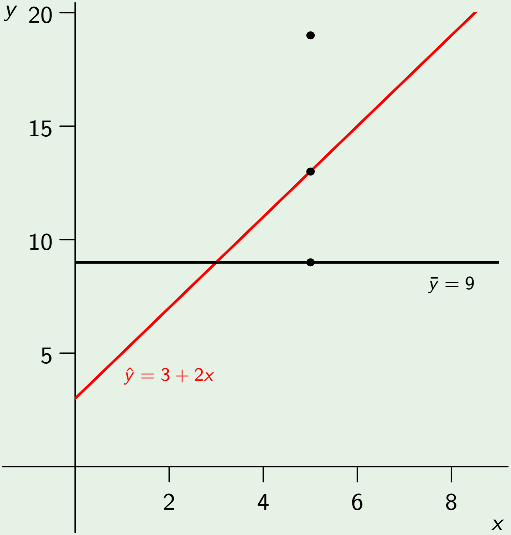
Example 2



We assume the following:

- There is sufficient evidence to support the claim of a linear correlation between x and y.
- The equation of the regression 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 $y = 19$.

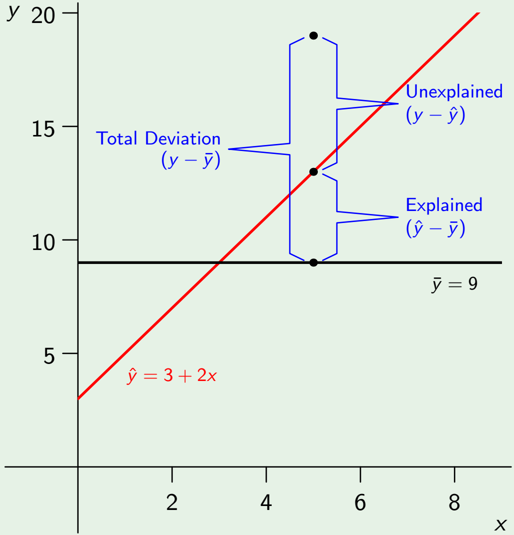
Example 2



We assume the following:

- There is sufficient evidence to support the claim of a linear correlation between x and y.
- The equation of the regression 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 substitution $x = 5$ into the regression equation yields $\hat{y} = 13$.

Example 2



We assume the following:

- There is sufficient evidence to support the claim of a linear correlation between x and y.
- The equation of the regression 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 substitution $x = 5$ into the regression equation yields $\hat{y} = 13$.

Definition

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

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

Definition

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

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

Note

We can either compute r^2 using the variation or we can square the linear correlation coefficient r.

Example 3

For the paired data in Data Set 16 comparing chocolate consumption and Nobel laureate rate has linear correlation coefficient $r = 0.801$.

Example 3

For the paired data in Data Set 16 comparing chocolate consumption and Nobel laureate rate has linear correlation coefficient $r = 0.801$.

The coefficient of determination is $r^2 = 0.641$.

Example 3

For the paired data in Data Set 16 comparing chocolate consumption and Nobel laureate rate has linear correlation coefficient $r = 0.801$.

The coefficient of determination is $r^2 = 0.641$.

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.

Example 3

For the paired data in Data Set 16 comparing chocolate consumption and Nobel laureate rate has linear correlation coefficient $r = 0.801$.

The coefficient of determination is $r^2 = 0.641$.

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.

Note

The other 35.8% might be explained by some other factors. But it is pretty silly to seriously think that chocolate consumption in a country is going to directly effect the country's rate of Nobel Laureates.