Regression

Colby Community College

Recall

Recall that a linear equation (a straight line) is one that can be written as

$$y = mx + b$$

where m is the slope ("rise over run") and b is the y-intercept.

Recall

Recall that a linear equation (a straight line) is one that can be written as

$$y = mx + b$$

where m is the slope ("rise over run") and b is the y-intercept.

Definition

Given a collection of paired sample data, the **regression line** (or **line of best fit**) is the straight line that "best" fits the scatter plot of the data. (We will discuss that "best" means later.)

Definition

The **regression equation** is

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the regression line. The regression equation expresses a relationship between x and $\hat{y}$.

Definition

The **regression equation** is

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the regression line. The regression equation expresses a relationship between x and $\hat{y}$.

Definition

We call x the **explanatory variable**, **predictor variable**, or **independent variable**.

Definition

The **regression equation** is

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the regression line. The regression equation expresses a relationship between x and $\hat{y}$.

Definition

We call x the **explanatory variable**, **predictor variable**, or **independent variable**.

Definition

We call y the **response variable**, or **dependent variable**.

Note

We don't use $y = mx + b$ because the format $y = b_0 + b_1 x$ can easily be expanded in include more variables:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \cdots$$

This is used when performing a multiple regression.

Requirements

The requirements for performing a regression are:

1. The sample of paired data is a random sample of quantitative data.

Requirements

The requirements for performing a regression are:

1. The sample of paired data is a random sample of quantitative data.
2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.

Requirements

The requirements for performing a regression are:

1. The sample of paired data is a random sample of quantitative data.
2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.
3. Outliers can have a strong effect on the regression equation, so remove any outliers if they are known errors.

Slope

The slope of the regression line is

$$b_1 = r \cdot \frac{s_y}{s_x}$$

where r is the linear correlation coefficient, s_y is the standard deviation of the y values, and s_x is the standard deviation of the x values.

Slope

The slope of the regression line is

$$b_1 = r \cdot \frac{s_y}{s_x}$$

where r is the linear correlation coefficient, s_y is the standard deviation of the y values, and s_x is the standard deviation of the x values.

y-intercept

The y-intercept of the regression line is

$$b_0 = \bar{y} - b_1\bar{x}$$

where $\bar{y}$ is the sample mean of y values and $\bar{x}$ is the sample mean of x values.

Slope

The slope of the regression line is

$$b_1 = r \cdot \frac{s_y}{s_x}$$

where r is the linear correlation coefficient, s_y is the standard deviation of the y values, and s_x is the standard deviation of the x values.

y-intercept

The y-intercept of the regression line is

$$b_0 = \bar{y} - b_1\bar{x}$$

where $\bar{y}$ is the sample mean of y values and $\bar{x}$ is the sample mean of x values.

Note

Technology will calculate both of these values for you.

Making Predictions

When making predictions, keep the following in mind:

Bad Model: If the regression equation does not appear to be useful for making predictions, don't use the regression equation.

Making Predictions

When making predictions, keep the following in mind:

Bad Model: If the regression equation does not appear to be useful for making predictions, don't use the regression equation.

- The best predicted value of a variable in this case is simply its sample mean.

Making Predictions

When making predictions, keep the following in mind:

Bad Model: If the regression equation does not appear to be useful for making predictions, don't use the regression equation.

- The best predicted value of a variable in this case is simply its sample mean.

Good Model: Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.

Making Predictions

When making predictions, keep the following in mind:

Bad Model: If the regression equation does not appear to be useful for making predictions, don't use the regression equation.

- The best predicted value of a variable in this case is simply its sample mean.

Good Model: Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.

Correlation: Use the regression equation for predictions only if the linear correlation coefficient r indicates that there is a linear correction between the two variables.

Making Predictions

When making predictions, keep the following in mind:

Bad Model: If the regression equation does not appear to be useful for making predictions, don't use the regression equation.

- The best predicted value of a variable in this case is simply its sample mean.

Good Model: Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.

Correlation: Use the regression equation for predictions only if the linear correlation coefficient r indicates that there is a linear correction between the two variables.

Scope: Use the regression line for predictions only if the data do no go much beyond the scope of available sample data.

Making Predictions

When making predictions, keep the following in mind:

Bad Model: If the regression equation does not appear to be useful for making predictions, don't use the regression equation.

- The best predicted value of a variable in this case is simply its sample mean.

Good Model: Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.

Correlation: Use the regression equation for predictions only if the linear correlation coefficient r indicates that there is a linear correction between the two variables.

Scope: Use the regression line for predictions only if the data do no go much beyond the scope of available sample data.

- Predicting too far beyond the scope of the available sample data is called **extrapolation** and can easily result in bad predictions.

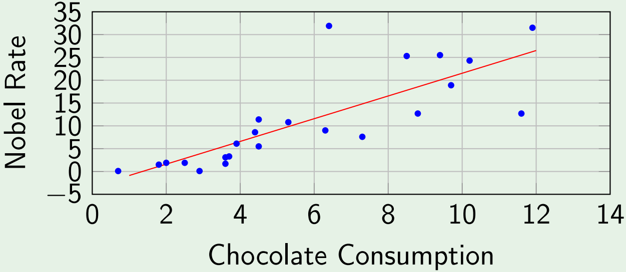
Example 1

Using the Data Set 16, we can compare a countries chocolate consumption to the number of Nobel laureates.

Example 1

Using the Data Set 16, we can compare a countries chocolate consumption to the number of Nobel laureates.

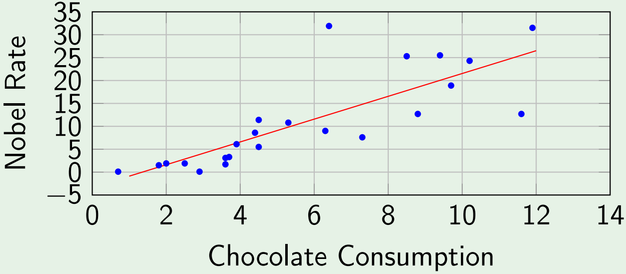
Technology give a regression line of $\hat{y} = -3.37 + 2.49x$ and the scatterplot:



Example 1

Using the Data Set 16, we can compare a countries chocolate consumption to the number of Nobel laureates.

Technology give a regression line of $\hat{y} = -3.37 + 2.49x$ and the scatterplot:



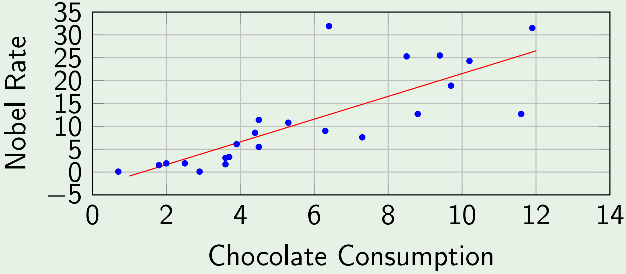
This is a good model, so we can approximate the Nobel Laureate rate for a country that consumes 10kg per capita:

$$\hat{y} = -3.37 + 2.49(10)$$

Example 1

Using the Data Set 16, we can compare a countries chocolate consumption to the number of Nobel laureates.

Technology give a regression line of $\hat{y} = -3.37 + 2.49x$ and the scatterplot:



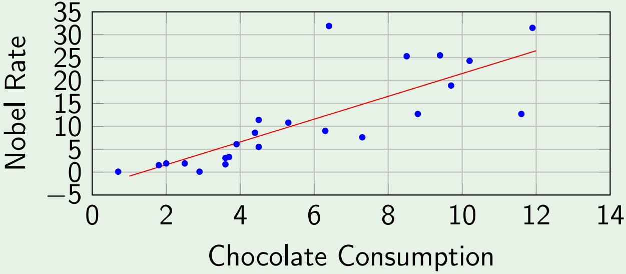
This is a good model, so we can approximate the Nobel Laureate rate for a country that consumes 10kg per capita:

$$\hat{y} = -3.37 + 2.49(10) = 21.5$$

Example 1

Using the Data Set 16, we can compare a countries chocolate consumption to the number of Nobel laureates.

Technology give a regression line of $\hat{y} = -3.37 + 2.49x$ and the scatterplot:



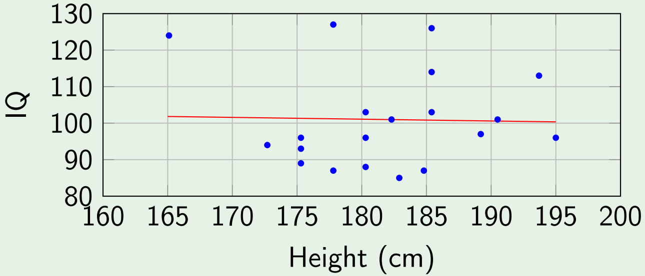
This is a good model, so we can approximate the Nobel Laureate rate for a country that consumes 10kg per capita:

$$\hat{y} = -3.37 + 2.49(10) = 21.5$$

So, we expect 21.5 Nobel Laureates per 10 million people.

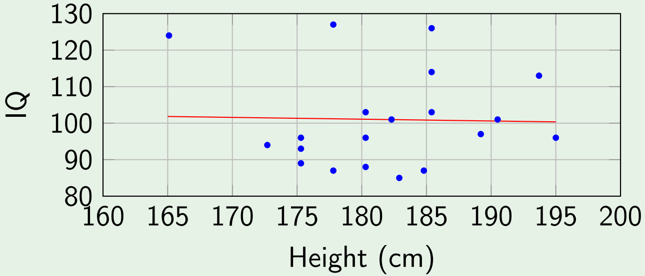
Example 2

The scatterplot shows sample data recording subject height and IQ score.



Example 2

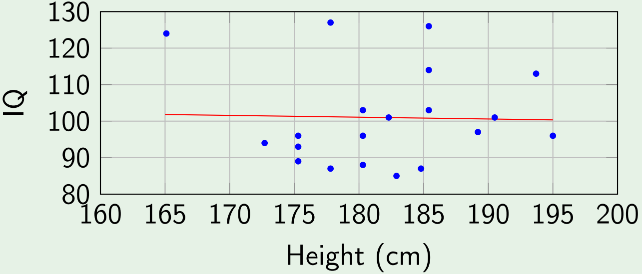
The scatterplot shows sample data recording subject height and IQ score.



The correlation is $r = -0.02731$ and has P-value 0.909.

Example 2

The scatterplot shows sample data recording subject height and IQ score.



The correlation is $r = -0.02731$ and has P-value 0.909.

This means the regression line is a bad model and should not be used to make predictions.