Least Squares Fitting
Wilfrid Laurier University

Terry Sturtevant

Wilfrid Laurier University

December 12, 2014

Overview

Overview

In this document, you'll learn:

Overview

In this document, you'll learn:

- how to determine the slope and y-intercept from straight line data

Overview

In this document, you'll learn:

- how to determine the slope and y-intercept from straight line data
- how to determine standard errors in the slope and y-intercept

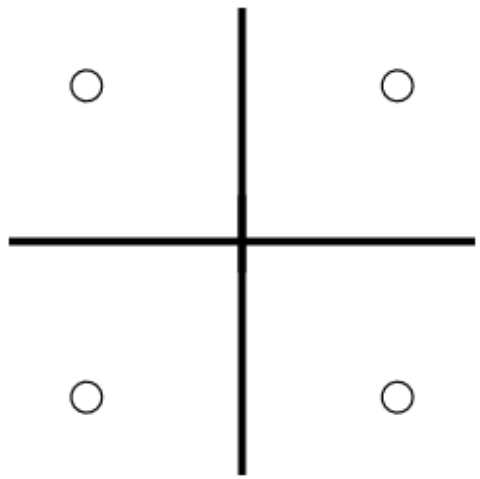
What does "line of best fit" mean?

What does "line of best fit" mean?

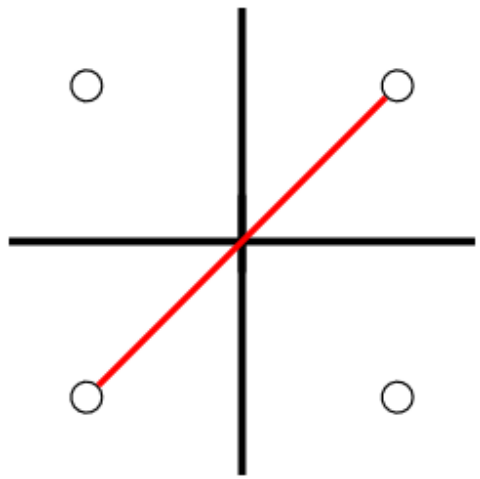
Unless a set of data *exactly* fits a curve, choosing *a* curve of "best fit" is somewhat arbitrary.

What does "line of best fit" mean?

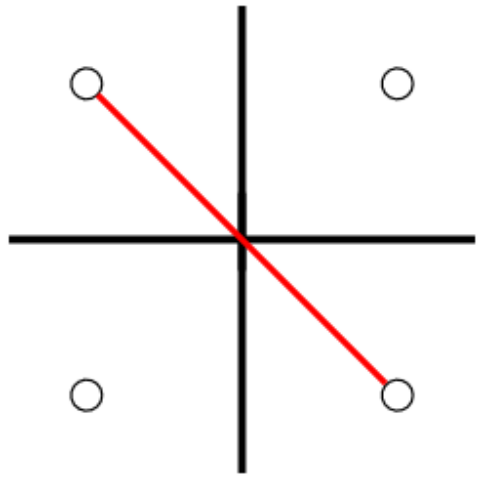
Unless a set of data *exactly* fits a curve, choosing *a* curve of "best fit" is somewhat arbitrary. (For example, consider 4 data points at (-1,1), (1,1), (1,-1) and (-1,-1). What line fits these points best?)



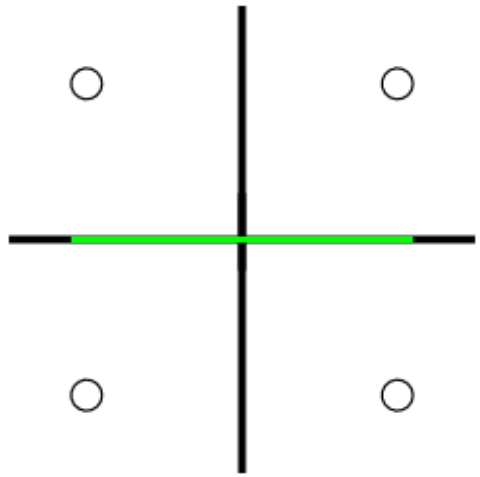
What line best fits these points?



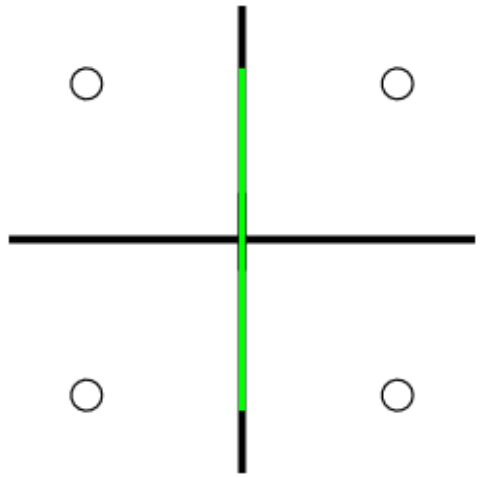
This line goes through two points exactly, and is equidistant from the others.



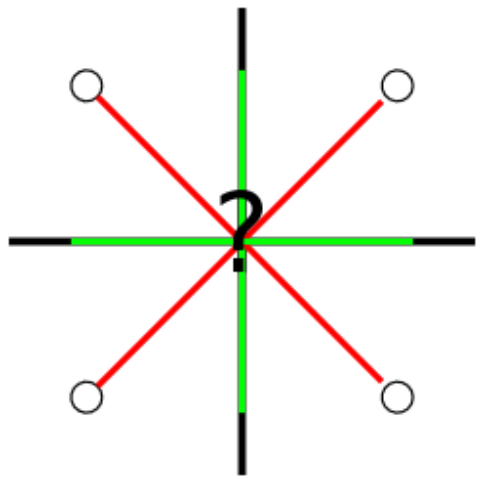
But then so does this one....



This line is equidistant from all of the points.



But so is this one.....



There is no clear "best" line through these four points.

Since there is no “universal” definition of what it means for a line to fit a set of points “best”, the least squares fit is used because it has certain mathematical properties that make it convenient.

Since there is no "universal" definition of what it means for a line to fit a set of points "best", the least squares fit is used because it has certain mathematical properties that make it convenient. Specifically:

Since there is no "universal" definition of what it means for a line to fit a set of points "best", the least squares fit is used because it has certain mathematical properties that make it convenient. Specifically:

- it always produces a solution

Since there is no "universal" definition of what it means for a line to fit a set of points "best", the least squares fit is used because it has certain mathematical properties that make it convenient. Specifically:

- it always produces a solution
- it can be updated for additional data points without complete recalculation

Least Squares Fitting

Least Squares Fitting

Least **S**quares **F**itting is a procedure for numerically determining the equation of a curve which "best approximates" the data being plotted.

Least Squares Fitting

Least **S**quares **F**itting is a procedure for numerically determining the equation of a curve which "best approximates" the data being plotted. If we wish to fit a straight line to data in the form

Least Squares Fitting

Least **S**quares **F**itting is a procedure for numerically determining the equation of a curve which "best approximates" the data being plotted. If we wish to fit a straight line to data in the form

$$y = mx + b$$

Least Squares Fitting

Least **S**quares **F**itting is a procedure for numerically determining the equation of a curve which "best approximates" the data being plotted. If we wish to fit a straight line to data in the form

$$y = mx + b$$

then the least squares fit gives values for b, the y-intercept, and m, the slope, as follows:

$$b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

$$b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

and

$$b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

and

$$m = \frac{N\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

You may notice that a particular quantity comes up a lot.

You may notice that a particular quantity comes up a lot. It is

You may notice that a particular quantity comes up a lot. It is

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2$$

You may notice that a particular quantity comes up a lot. It is

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2$$

It only takes a couple of lines of algebra to show that this equals

You may notice that a particular quantity comes up a lot. It is

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2$$

It only takes a couple of lines of algebra to show that this equals

$$N\left(N-1\right)\sigma_x{}^2$$

You may notice that a particular quantity comes up a lot. It is

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2$$

It only takes a couple of lines of algebra to show that this equals

$$N\left(N-1\right){\sigma_x}^2$$

where ${\sigma_x}^2$ is the sample standard deviation of the x values.

(Note: You do not need to calculate uncertainties for m and b during least squares fit calculations like this.

(Note: You do not need to calculate uncertainties for m and b during least squares fit calculations like this. Uncertainties in m and b will be dealt with later.)

One important concept which will come up later is that of **degrees of freedom**,

One important concept which will come up later is that of **degrees of freedom**, which is simply the number which is the difference between the number of data points, (N above), and the number of parameters being determined by the fit, (2 for a straight line).

One important concept which will come up later is that of **degrees of freedom**, which is simply the number which is the difference between the number of data points, (N above), and the number of parameters being determined by the fit, (2 for a straight line). Thus, for a linear fit, the number of degrees of freedom, ν, is given by:

One important concept which will come up later is that of **degrees of freedom**, which is simply the number which is the difference between the number of data points, (N above), and the number of parameters being determined by the fit, (2 for a straight line). Thus, for a linear fit, the number of degrees of freedom, ν, is given by:

$$\nu = N - 2$$

Correlation Coefficient

Correlation Coefficient

The following equation gives the square of the **Pearson product-moment correlation coefficient**,

Correlation Coefficient

The following equation gives the square of the **Pearson product-moment correlation coefficient**, which we will refer to simply as the *correlation coefficient*.

Correlation Coefficient

The following equation gives the square of the **Pearson product-moment correlation coefficient**, which we will refer to simply as the *correlation coefficient*.

$$R^2 = \frac{(N\sum x_i y_i - (\sum x_i)(\sum y_i))^2}{(N\sum {x_i}^2 - (\sum x_i)^2)\left(N\sum {y_i}^2 - (\sum y_i)^2\right)}$$

Correlation Coefficient

The following equation gives the square of the **Pearson product-moment correlation coefficient**, which we will refer to simply as the *correlation coefficient*.

$$R^2 = \frac{(N\sum x_i y_i - (\sum x_i)(\sum y_i))^2}{(N\sum {x_i}^2 - (\sum x_i)^2)\left(N\sum {y_i}^2 - (\sum y_i)^2\right)}$$

As long as we're dealing with a *linear* fit, this is the quantity that would commonly be used.

The correlation coefficient, R, is a number which has a value between -1 and $+1$, where

The correlation coefficient, R, is a number which has a value between -1 and $+1$, where

- a value of -1 indicates a perfect *negative* correlation,

The correlation coefficient, R, is a number which has a value between -1 and $+1$, where

- a value of -1 indicates a perfect *negative* correlation,
- a value of $+1$ indicates a perfect *positive* correlation,

The correlation coefficient, R, is a number which has a value between -1 and $+1$, where

- a value of -1 indicates a perfect *negative* correlation,
- a value of $+1$ indicates a perfect *positive* correlation,
- a value of zero indicates *no* correlation.

The correlation coefficient, R, is a number which has a value between -1 and $+1$, where

- a value of -1 indicates a perfect *negative* correlation,
- a value of $+1$ indicates a perfect *positive* correlation,
- a value of zero indicates *no* correlation.

Thus R^2 is a value between zero and 1 indicating just the *strength* of a correlation.

The correlation coefficient, R, is a number which has a value between -1 and $+1$, where

- a value of -1 indicates a perfect *negative* correlation,
- a value of $+1$ indicates a perfect *positive* correlation,
- a value of zero indicates *no* correlation.

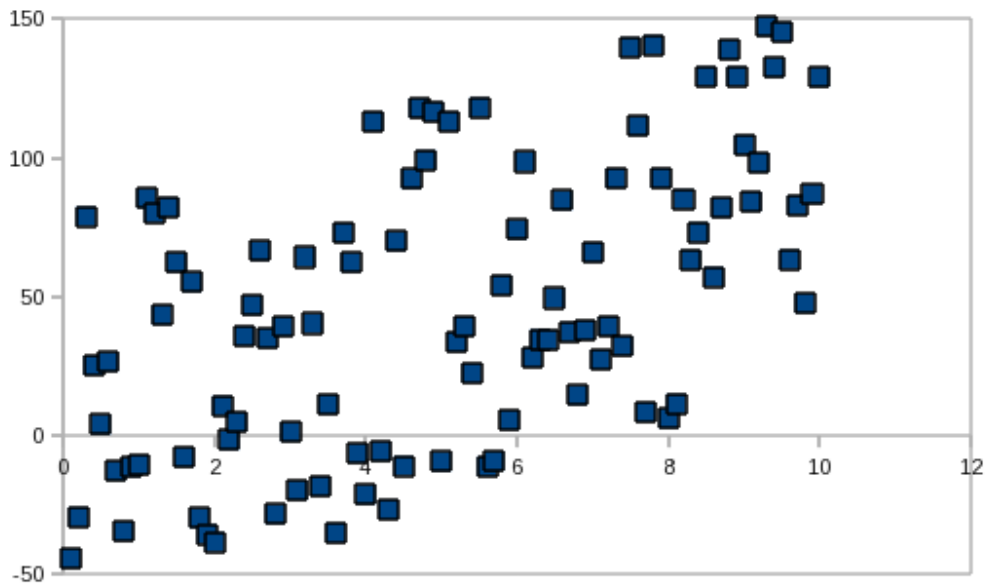
Thus R^2 is a value between zero and 1 indicating just the *strength* of a correlation. The closer R^2 is to one, the stronger the correlation between two variables.

The correlation coefficient, R, is a number which has a value between -1 and $+1$, where

- a value of -1 indicates a perfect *negative* correlation,
- a value of $+1$ indicates a perfect *positive* correlation,
- a value of zero indicates *no* correlation.

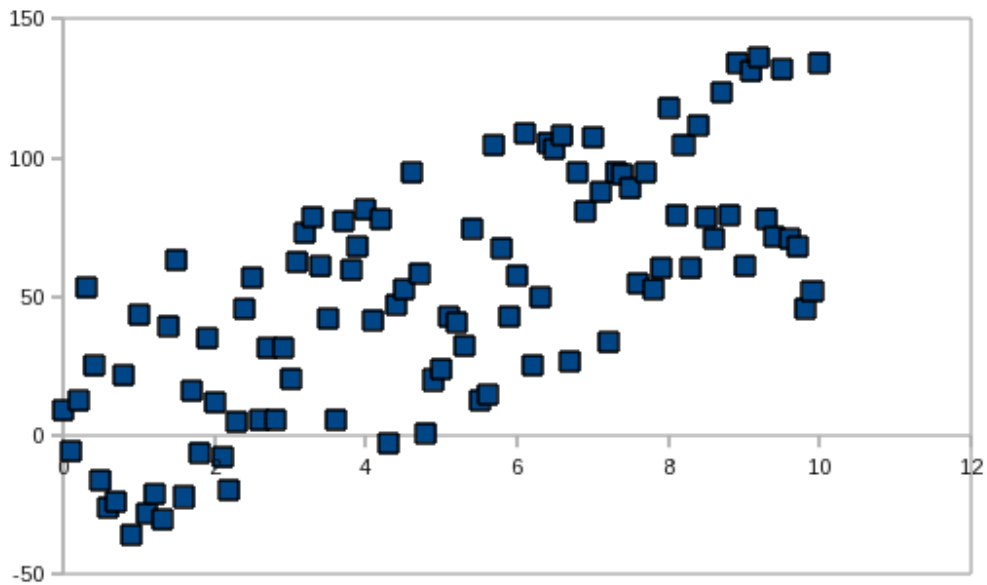
Thus R^2 is a value between zero and 1 indicating just the *strength* of a correlation. The closer R^2 is to one, the stronger the correlation between two variables. To put it another way, the closer it is to one the better one variable can be used as a *predictor* of the other.

Visualization of R^2



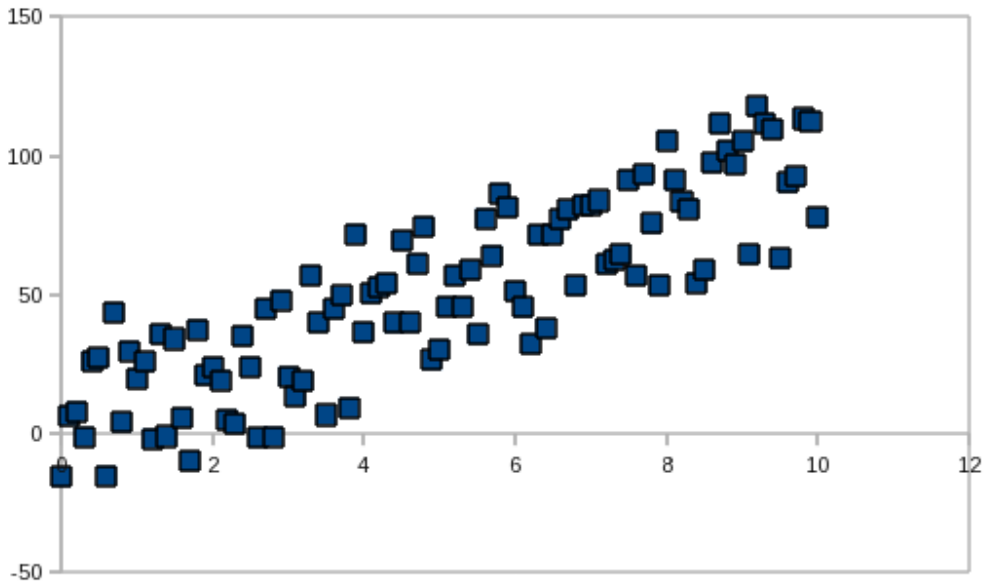
$R^2 \approx 0.3$

Visualization of R^2



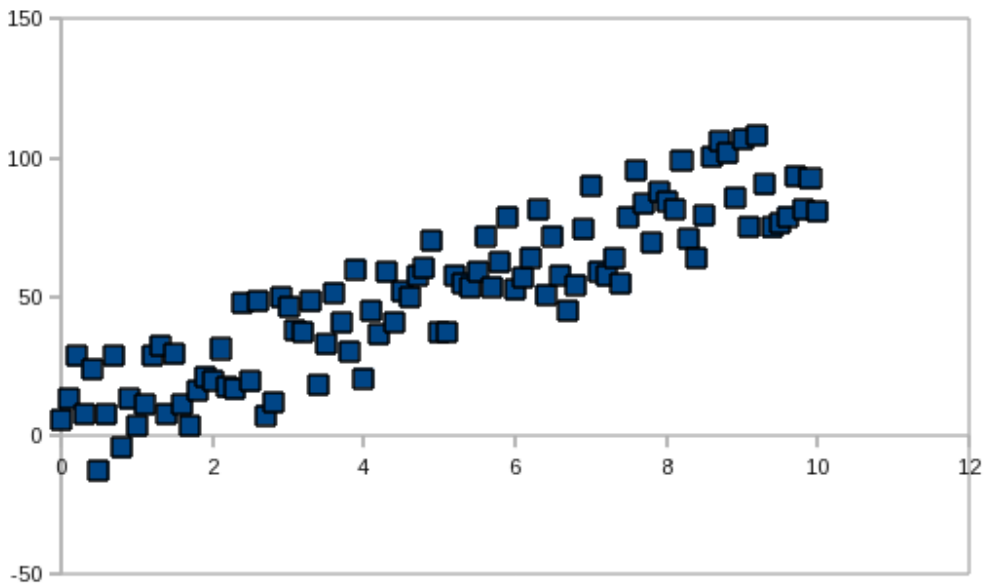
$R^2 \approx 0.5$

Visualization of R^2



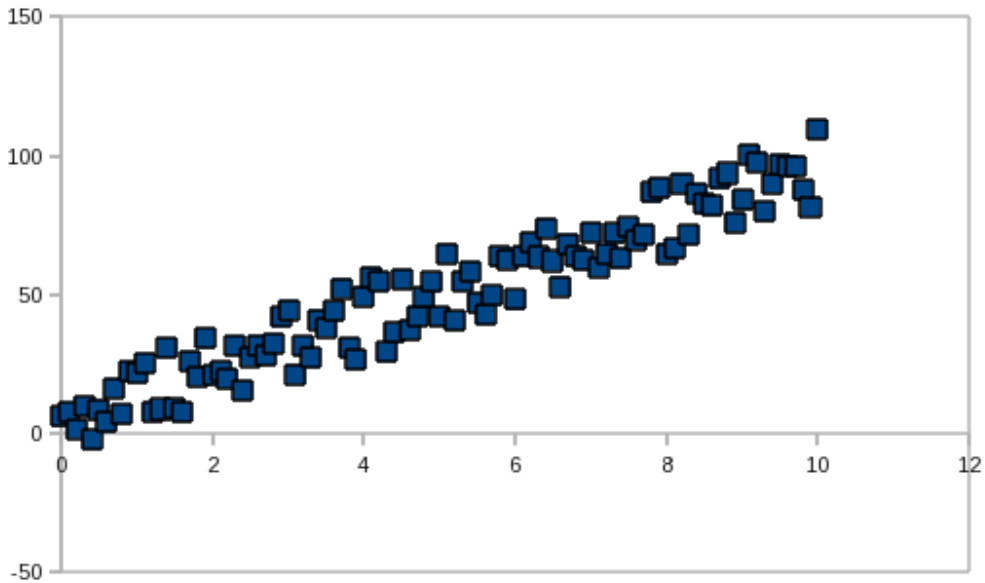
$R^2 \approx 0.7$

Visualization of R^2



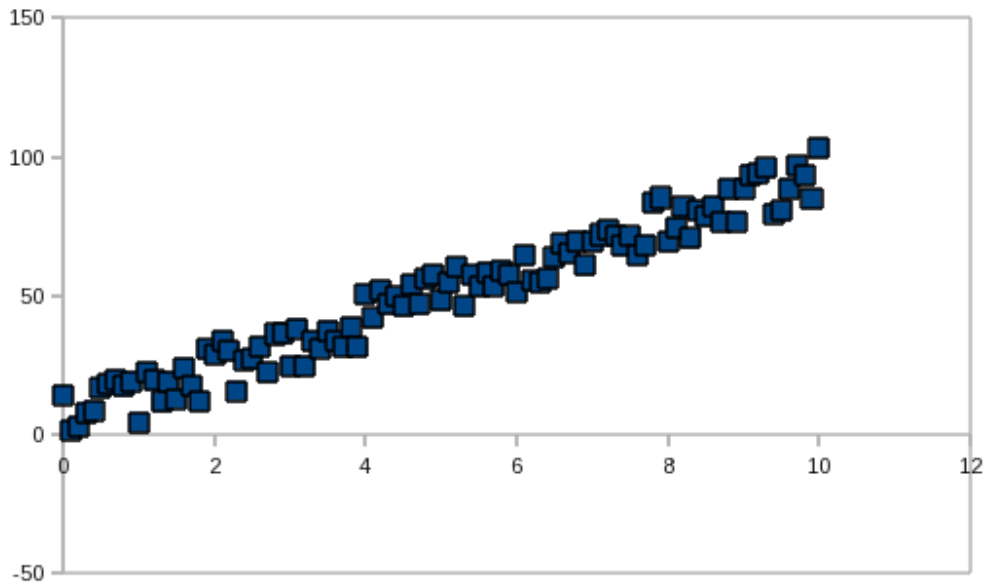
$R^2 \approx 0.8$

Visualization of R^2



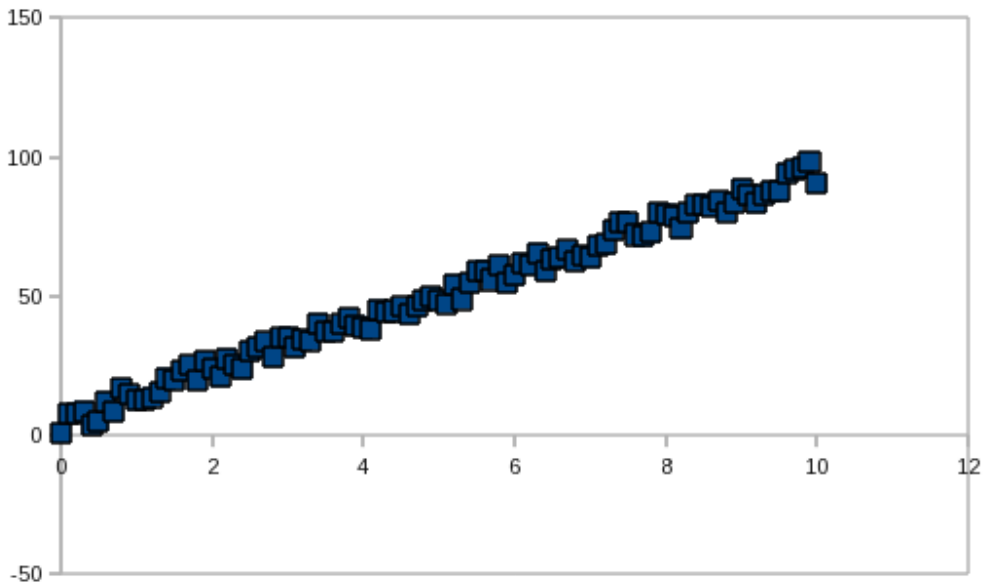
$R^2 \approx 0.9$

Visualization of R^2



$R^2 \approx 0.95$

Visualization of R^2



$R^2 \approx 0.99$

Once the values for the slope and intercept are determined, the sum of squares error, S is computed.

Once the values for the slope and intercept are determined, the sum of squares error, S is computed. For the linear case, S can be shown to have a value of

Once the values for the slope and intercept are determined, the sum of squares error, S is computed. For the linear case, S can be shown to have a value of

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$

Once the values for the slope and intercept are determined, the sum of squares error, S is computed. For the linear case, S can be shown to have a value of

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$

In order to estimate the uncertainty in each parameter, the standard deviation σ is computed from

Once the values for the slope and intercept are determined, the sum of squares error, S is computed. For the linear case, S can be shown to have a value of

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$

In order to estimate the uncertainty in each parameter, the standard deviation σ is computed from

$$\sigma = \sqrt{\frac{S}{N-2}}$$

Once the values for the slope and intercept are determined, the sum of squares error, S is computed. For the linear case, S can be shown to have a value of

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$

In order to estimate the uncertainty in each parameter, the standard deviation σ is computed from

$$\sigma = \sqrt{\frac{S}{N-2}}$$

where $N - 2$ is the number of degrees of freedom mentioned earlier.

Once the values for the slope and intercept are determined, the sum of squares error, S is computed. For the linear case, S can be shown to have a value of

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$

In order to estimate the uncertainty in each parameter, the standard deviation σ is computed from

$$\sigma = \sqrt{\frac{S}{N-2}}$$

where $N - 2$ is the number of degrees of freedom mentioned earlier. (Often the symbol ν is used for degrees of freedom.)

The **standard error** in the intercept is

The **standard error** in the intercept is

$$\sigma_b = \sigma \sqrt{\frac{\sum x_i^2}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}}$$

The **standard error** in the intercept is

$$\sigma_b = \sigma \sqrt{\frac{\sum x_i^2}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}}$$

and the standard error in the slope is

The **standard error** in the intercept is

$$\sigma_b = \sigma \sqrt{\frac{\sum x_i^2}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}}$$

and the standard error in the slope is

$$\sigma_m = \sigma \sqrt{\frac{N}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}}$$

Following is a calculation of the least squares fit and the standard error of the slope and intercept for some test data.

i	x_i	x_i^2	$x_i y_i$	y_i	y_i^2
1	0.1			3	
N	$\sum x_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum y_i$	$\sum y_i^2$
1	0.1			3	

Sample Least Squares Fit Data

Following is a calculation of the least squares fit and the standard error of the slope and intercept for some test data.

i	x_i	x_i^2	$x_i y_i$	y_i	y_i^2
1	0.1	0.01		3	
N	$\sum x_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum y_i$	$\sum y_i^2$
1	0.1	0.01		3	

Sample Least Squares Fit Data

Following is a calculation of the least squares fit and the standard error of the slope and intercept for some test data.

i	x_i	x_i^2	$x_i y_i$	y_i	y_i^2
1	0.1	0.01	0.3	3	
N	$\sum x_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum y_i$	$\sum y_i^2$
1	0.1	0.01	0.3	3	

Sample Least Squares Fit Data

Following is a calculation of the least squares fit and the standard error of the slope and intercept for some test data.

i	x_i	x_i^2	$x_i y_i$	y_i	y_i^2
1	0.1	0.01	0.3	3	9
N	$\sum x_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum y_i$	$\sum y_i^2$
1	0.1	0.01	0.3	3	9

Sample Least Squares Fit Data

Following is a calculation of the least squares fit and the standard error of the slope and intercept for some test data.

i	x_i	x_i^2	$x_i y_i$	y_i	y_i^2
1	0.1	0.01	0.3	3	9
2	0.2	0.04	0.8	4	16
N	$\sum x_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum y_i$	$\sum y_i^2$
2	0.3	0.05	1.1	7	25

Sample Least Squares Fit Data

Following is a calculation of the least squares fit and the standard error of the slope and intercept for some test data.

i	x_i	x_i^2	$x_i y_i$	y_i	y_i^2
1	0.1	0.01	0.3	3	9
2	0.2	0.04	0.8	4	16
3	0.3	0.09	1.2	4	16
N	$\sum x_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum y_i$	$\sum y_i^2$
3	0.6	0.14	2.3	11	41

Sample Least Squares Fit Data

Following is a calculation of the least squares fit and the standard error of the slope and intercept for some test data.

i	x_i	x_i^2	$x_i y_i$	y_i	y_i^2
1	0.1	0.01	0.3	3	9
2	0.2	0.04	0.8	4	16
3	0.3	0.09	1.2	4	16
4	0.4	0.16	2.0	5	25
N	$\sum x_i$	$\sum x_i^2$	$\sum x_i y_i$	$\sum y_i$	$\sum y_i^2$
4	1.0	0.3	4.3	16	66

Sample Least Squares Fit Data

Calculation of b and m

Calculation of b and m

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2$$

Calculation of b and m

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 = (4)(0.3) - (1)^2$$

Calculation of b and m

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 = (4)(0.3) - (1)^2 = 0.2$$

Calculation of b and m

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 = (4)(0.3) - (1)^2 = 0.2$$

$$b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

Calculation of b and m

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 = (4)(0.3) - (1)^2 = 0.2$$

$$b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$
$$= \frac{(16)(0.3) - (1)(4.3)}{0.2}$$

Calculation of b and m

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 = (4)(0.3) - (1)^2 = 0.2$$

$$b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$
$$= \frac{(16)(0.3) - (1)(4.3)}{0.2} = 2.5$$

Calculation of b and m

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 = (4)(0.3) - (1)^2 = 0.2$$

$$b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$
$$= \frac{(16)(0.3) - (1)(4.3)}{0.2} = 2.5$$

$$m = \frac{N\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$

Calculation of b and m

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 = (4)(0.3) - (1)^2 = 0.2$$

$$b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} = \frac{(16)(0.3) - (1)(4.3)}{0.2} = 2.5$$

$$m = \frac{N\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} = \frac{(4)(4.3) - (1)(16)}{0.2}$$

Calculation of b and m

$$N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2 = (4)(0.3) - (1)^2 = 0.2$$

$$b = \frac{\left(\sum y_i\right)\left(\sum x_i^2\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}$$
$$= \frac{(16)(0.3) - (1)(4.3)}{0.2} = 2.5$$

$$m = \frac{N\left(\sum x_i y_i\right) - \left(\sum x_i\right)\left(\sum y_i\right)}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2} = \frac{(4)(4.3) - (1)(16)}{0.2} = 6.0$$

Calculation of S and σ

Calculation of S and σ

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$

Calculation of S and σ

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$
$$= (66) - (6)(4.3) - (2.5)(16)$$

Calculation of S and σ

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$
$$= (66) - (6)(4.3) - (2.5)(16) = 0.2$$

Calculation of S and σ

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$
$$= (66) - (6)(4.3) - (2.5)(16) = 0.2$$

$$\sigma = \sqrt{\frac{S}{N-2}}$$

Calculation of S and σ

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$
$$= (66) - (6)(4.3) - (2.5)(16) = 0.2$$

$$\sigma = \sqrt{\frac{S}{N-2}} = \sqrt{\frac{0.2}{4-2}}$$

Calculation of S and σ

$$S = \sum y_i^2 - m \sum x_i y_i - b \sum y_i$$
$$= (66) - (6)(4.3) - (2.5)(16) = 0.2$$

$$\sigma = \sqrt{\frac{S}{N-2}} = \sqrt{\frac{0.2}{4-2}} = 0.316228$$

Calculation of σ_b and σ_m

Calculation of σ_b and σ_m

$$\sigma_b = \sigma \sqrt{\frac{\sum x_i^2}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}}$$

Calculation of σ_b and σ_m

$$\sigma_b = \sigma \sqrt{\frac{\sum x_i^2}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}} = (0.316228)\sqrt{\frac{0.3}{0.2}}$$

Calculation of σ_b and σ_m

$$\sigma_b = \sigma \sqrt{\frac{\sum x_i^2}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}} = (0.316228)\sqrt{\frac{0.3}{0.2}} = (0.3878298)$$

Calculation of σ_b and σ_m

$$\sigma_b = \sigma \sqrt{\frac{\sum x_i^2}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}} = (0.316228)\sqrt{\frac{0.3}{0.2}} = (0.3878298)$$

$$\sigma_m = \sigma \sqrt{\frac{N}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}}$$

Calculation of σ_b and σ_m

$$\sigma_b = \sigma \sqrt{\frac{\sum x_i^2}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}} = (0.316228)\sqrt{\frac{0.3}{0.2}} = (0.3878298)$$

$$\sigma_m = \sigma \sqrt{\frac{N}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}} = (0.316228)\sqrt{\frac{4}{0.2}}$$

Calculation of σ_b and σ_m

$$\sigma_b = \sigma \sqrt{\frac{\sum x_i^2}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}} = (0.316228)\sqrt{\frac{0.3}{0.2}} = (0.3878298)$$

$$\sigma_m = \sigma \sqrt{\frac{N}{N\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}} = (0.316228)\sqrt{\frac{4}{0.2}} = (1.414214)$$

Recap

Recap

1. Least squares fitting allows you to find the slope and y-intercept of a straight line graph numerically (i.e. with a computer).

Recap

1. Least squares fitting allows you to find the slope and y-intercept of a straight line graph numerically (i.e. with a computer).
2. The standard errors in the slope and y-intercept will be used to determine the uncertainties in graphical results.

Recap

1. Least squares fitting allows you to find the slope and y-intercept of a straight line graph numerically (i.e. with a computer).
2. The standard errors in the slope and y-intercept will be used to determine the uncertainties in graphical results.
3. The correlation coefficient, R^2, is a measure of how closely data fit a line;

Recap

1. Least squares fitting allows you to find the slope and y-intercept of a straight line graph numerically (i.e. with a computer).
2. The standard errors in the slope and y-intercept will be used to determine the uncertainties in graphical results.
3. The correlation coefficient, R^2, is a measure of how closely data fit a line;

   the closer R^2 is to 1, the better the fit.