

Statistics with Spa OWS

Lecture 9

Julia Schroeder

Outline

- Linear models

Linear models

- Most important concepts
- Can model many questions (including previous t-test)
- If you understand linear models, everything else will be easy as a breeze!
- Aim to fit models to data

Fitting models to data



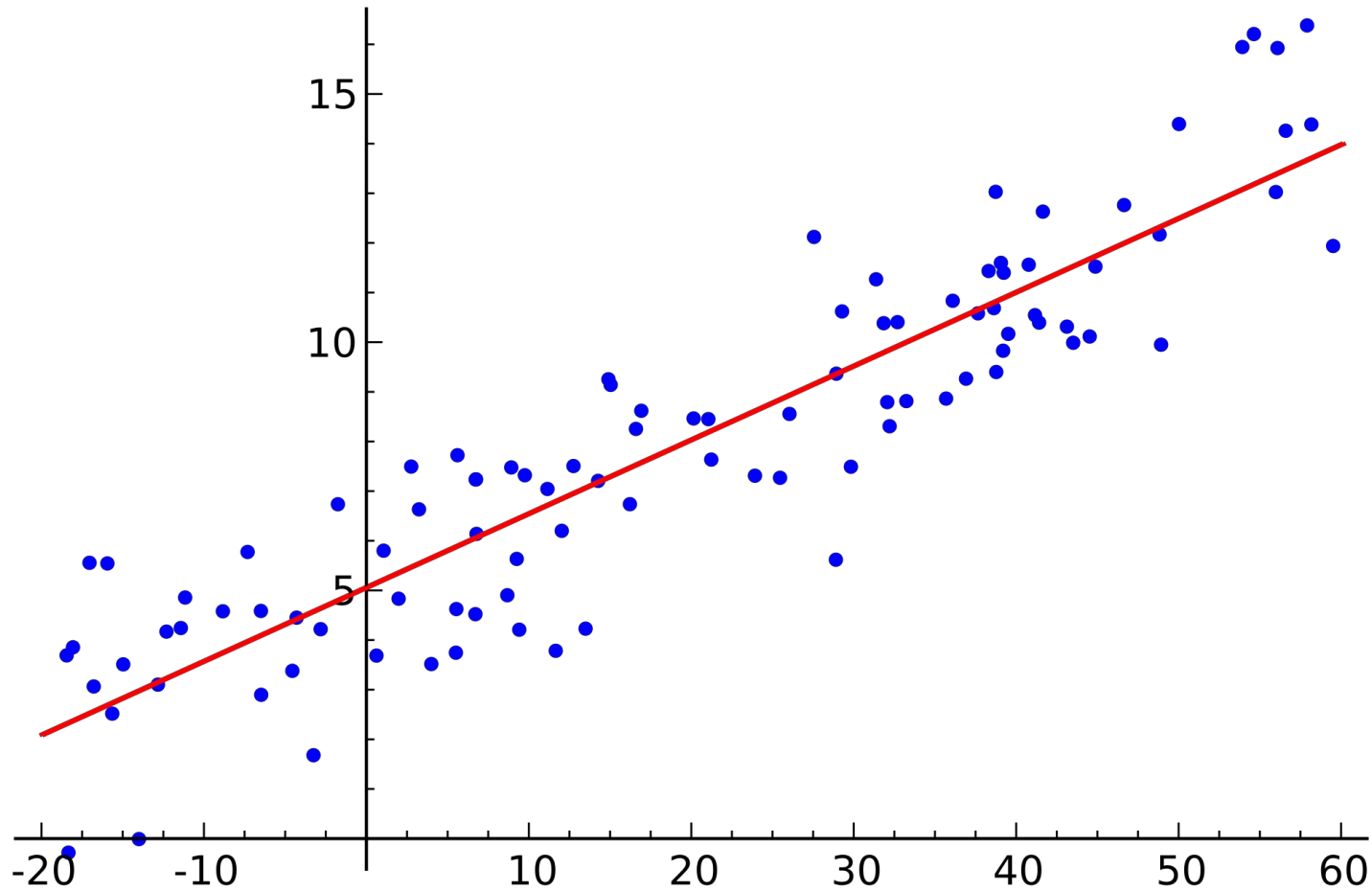
Fitting models to data



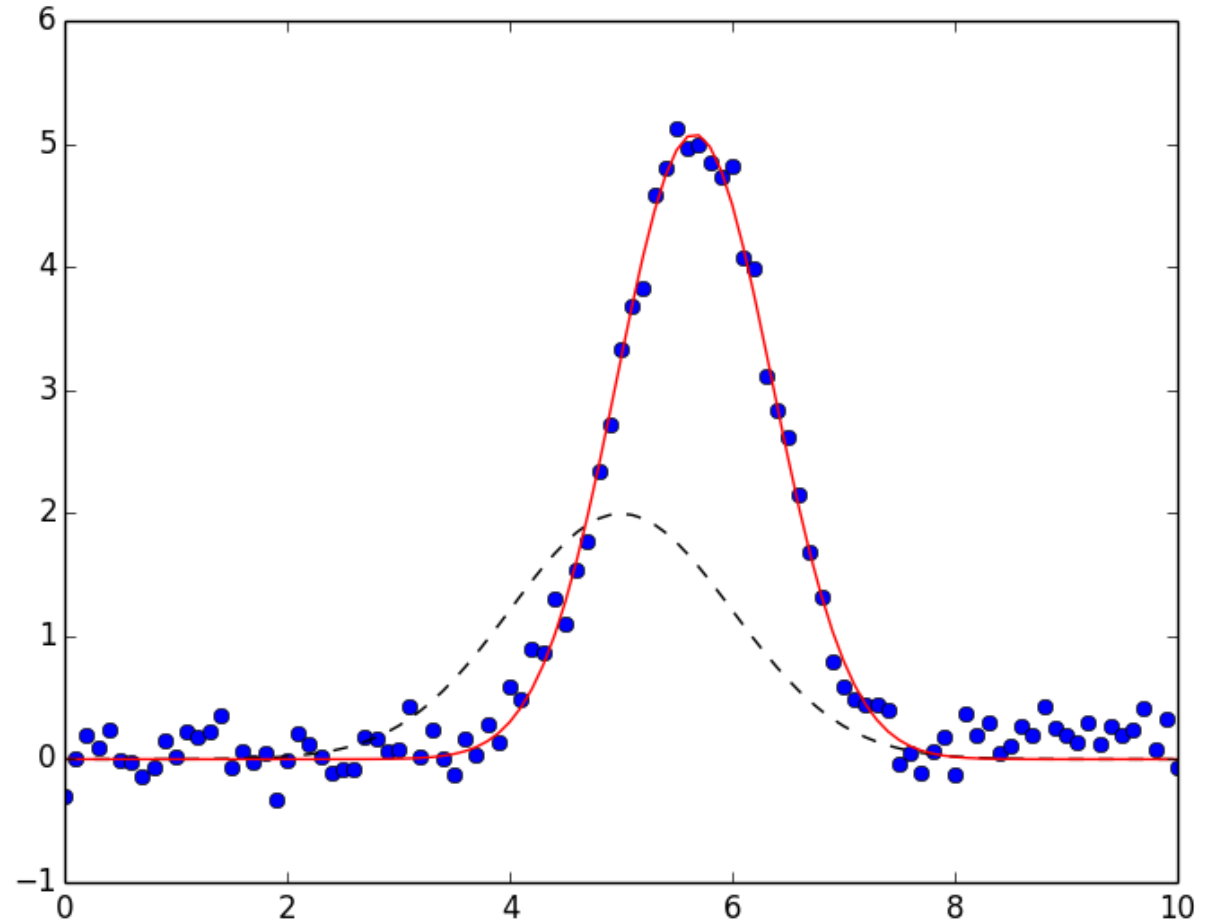
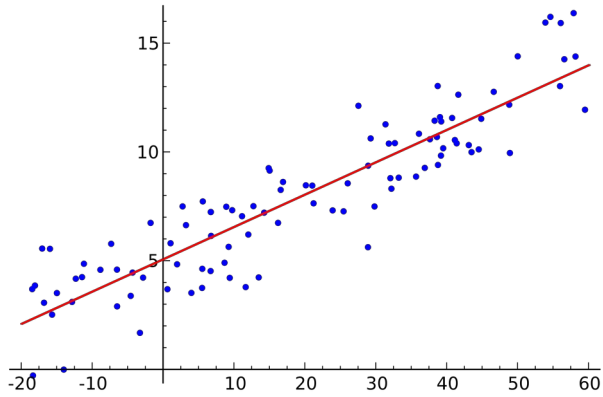
?



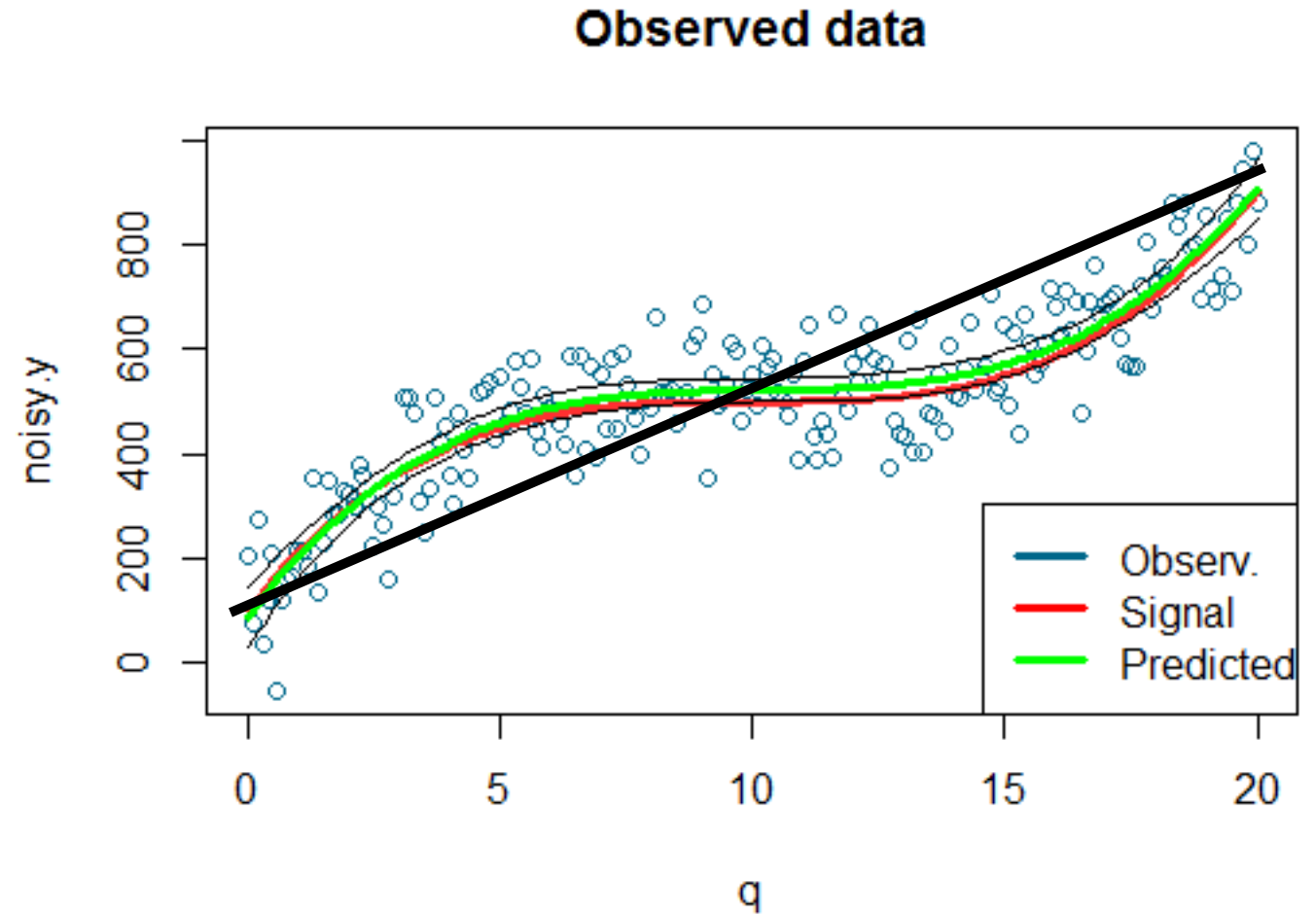
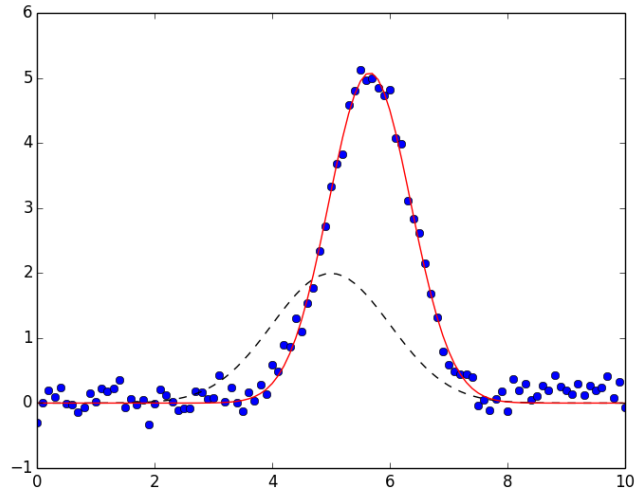
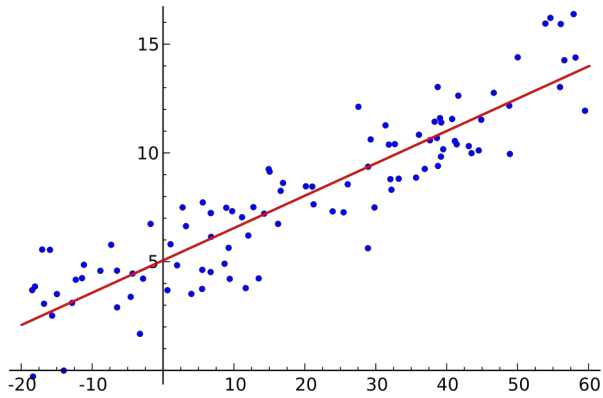
Fitting models to data



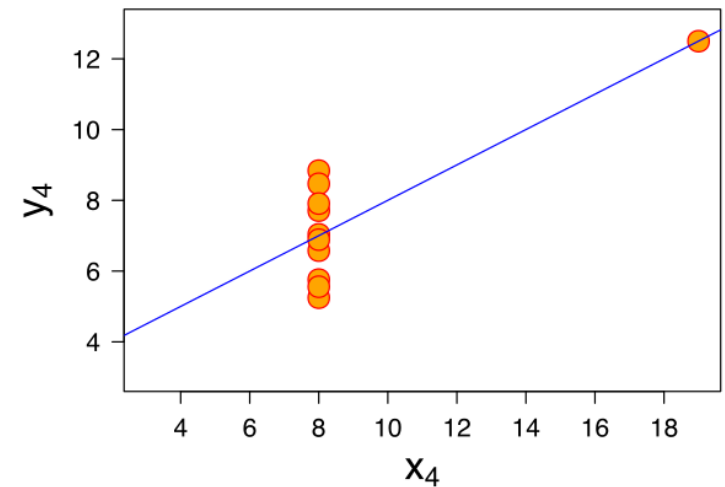
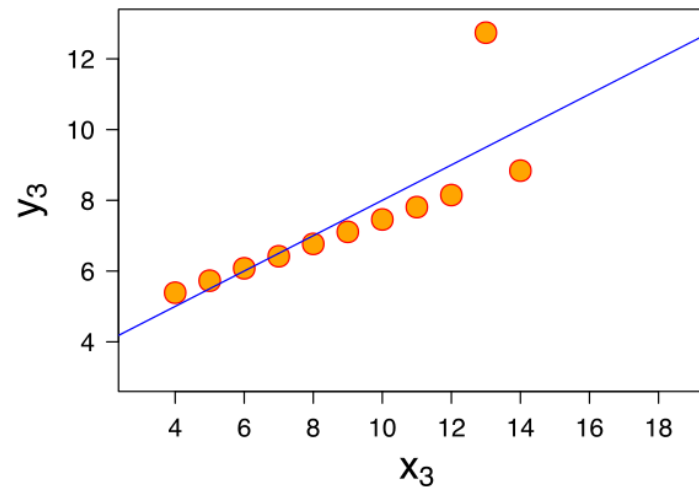
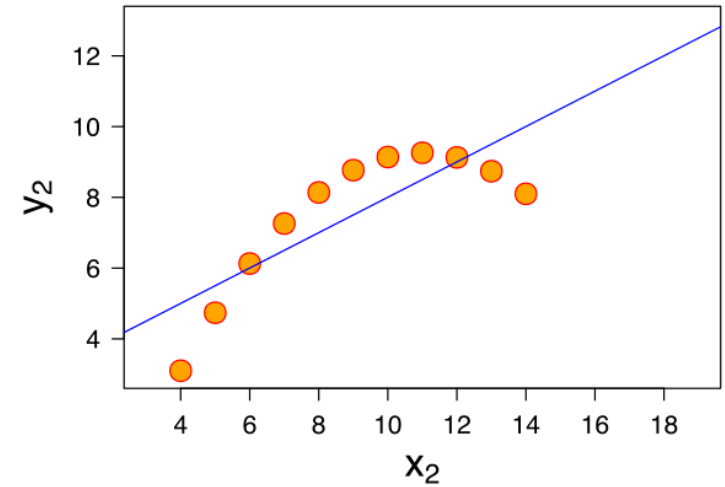
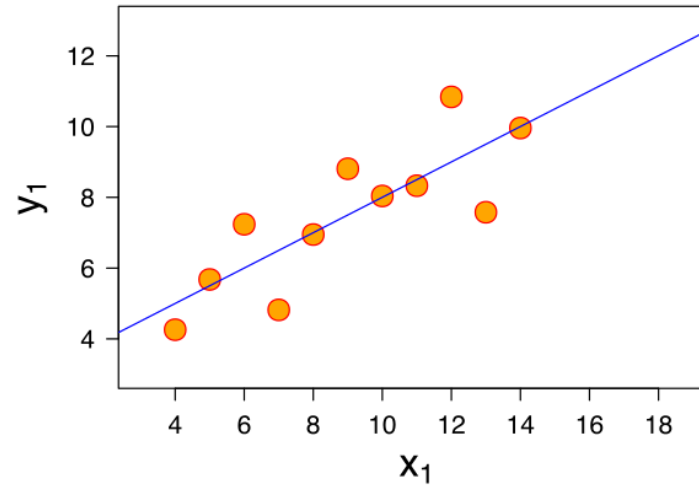
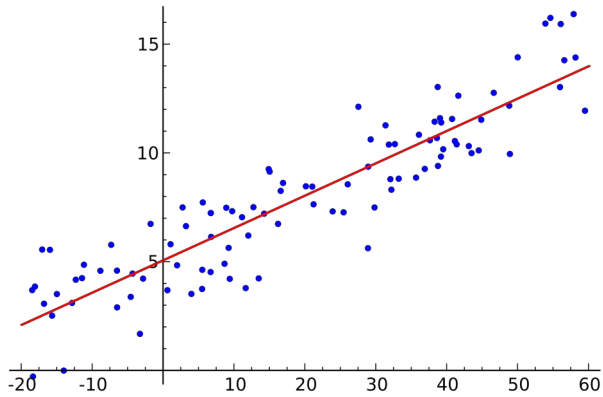
Fitting models to data



Fitting models to data



Fitting models to data



Linear models

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Linear models

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

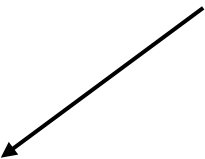


y
5
3
6
10
4

Data. Response variable,
e.g. sparrow body mass.

Linear models

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

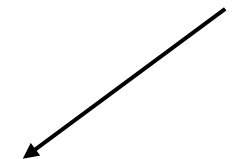


y	i
5	1
3	2
6	3
10	4
4	5

Data. Response variable. Observation 1, 2, 3, etc.
e.g. sparrow body mass.

Linear models

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$



y	i
5	1
3	2
6	3
10	4
4	5

Data. Response variable. Observation 1, 2, 3, etc.
e.g. sparrow body mass.



x	i
3	1
1	2
4	3
8	4
2	5

Data. Explanatory variable.
e.g. sparrow tarsus length.

Linear models

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$



	i
?	1
?	2
?	3
?	4
?	5

Linear models

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$



$b_1 = ?$

	i
?	1
?	2
?	3
?	4
?	5

Linear models

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$



$b_0 = ?$



$b_1 = ?$

?	i
?	1
?	2
?	3
?	4
?	5

Linear models

- Note difference in variable format
- Some are vectors, others are single values!

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

↙

y	i
5	1
3	2
6	3
10	4
4	5

↓

$b_0 = ?$

↓

$b_1 = ?$

↘

x	i		i
3	1	?	1
1	2	?	2
4	3	?	3
8	4	?	4
2	5	?	5

Linear models

- Note difference in variable format
- Some are vectors, others are single values!
- We aim to estimate b_0 and b_1
- We will get from the results

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Linear models

- Note difference in variable format
- Some are vectors, others are single values!
- We aim to estimate b_0 and b_1 *Parameter estimates*
- We will get ε_i from the results *Error, or residuals*

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Let's give this a try...

- Let's plot this

y
5
4
7
9
3

x
3
1
4
8
2

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

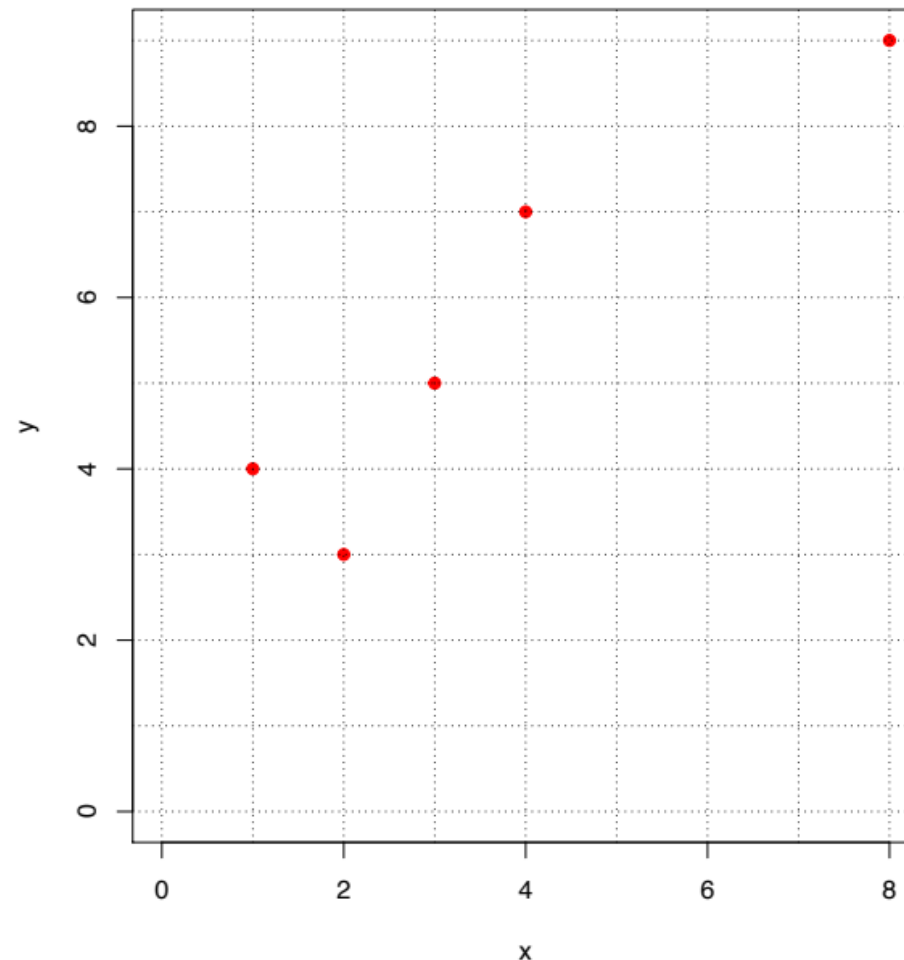
Let's give this a try...

- Let's plot this

y
5
4
7
9
3

x
3
1
4
8
2

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$



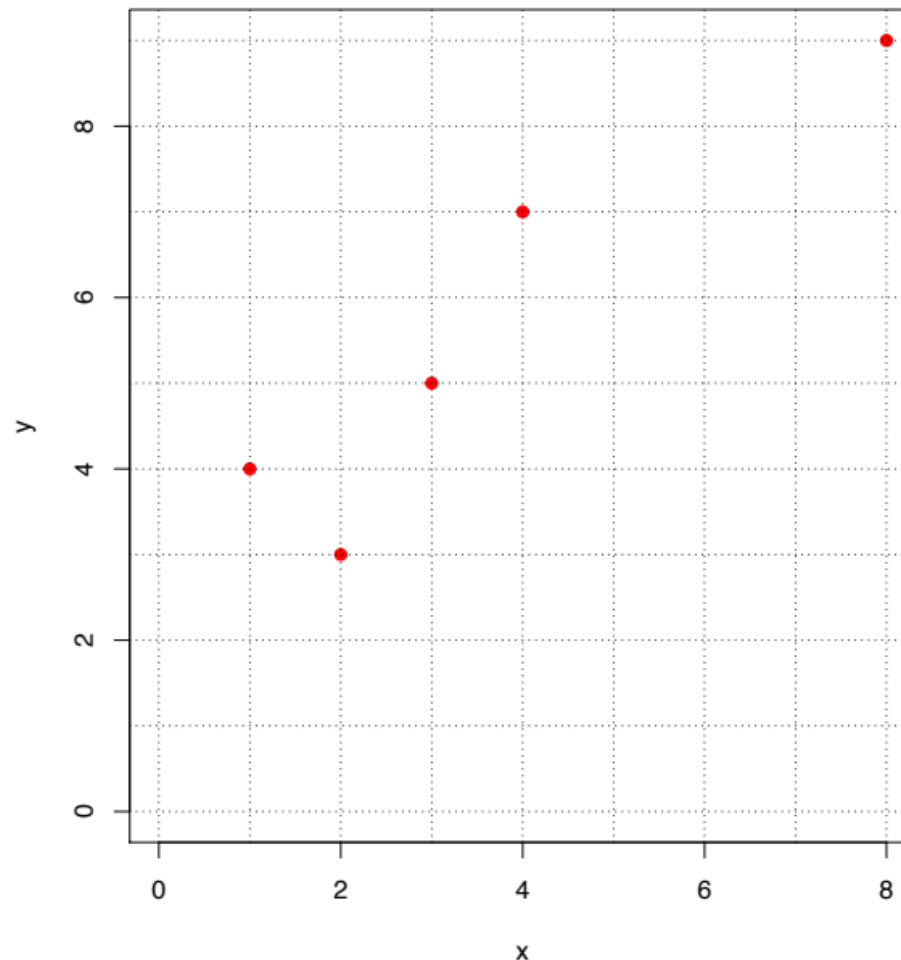
Let's give this a try...

- Let's plot this
- Now we “guesstimate” the line

y
5
4
7
9
3

x
3
1
4
8
2

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$



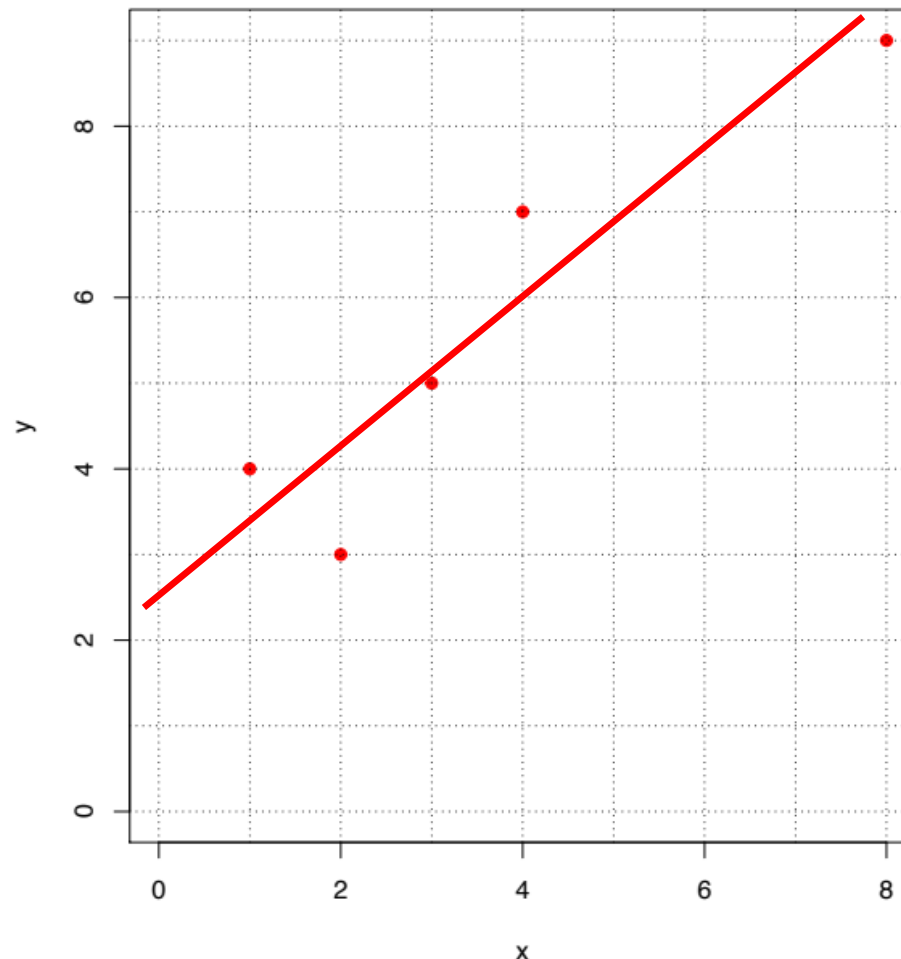
Let's give this a try...

- Let's plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1

y
5
4
7
9
3

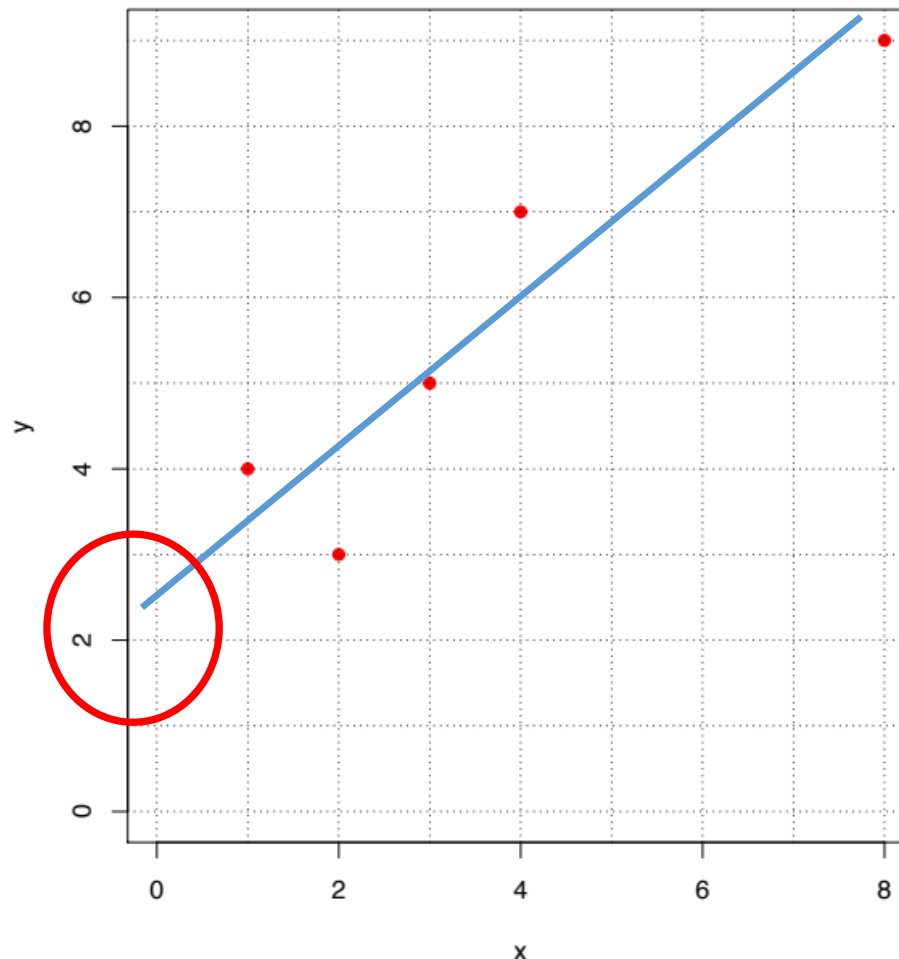
x
3
1
4
8
2

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$



Let's give this a try...

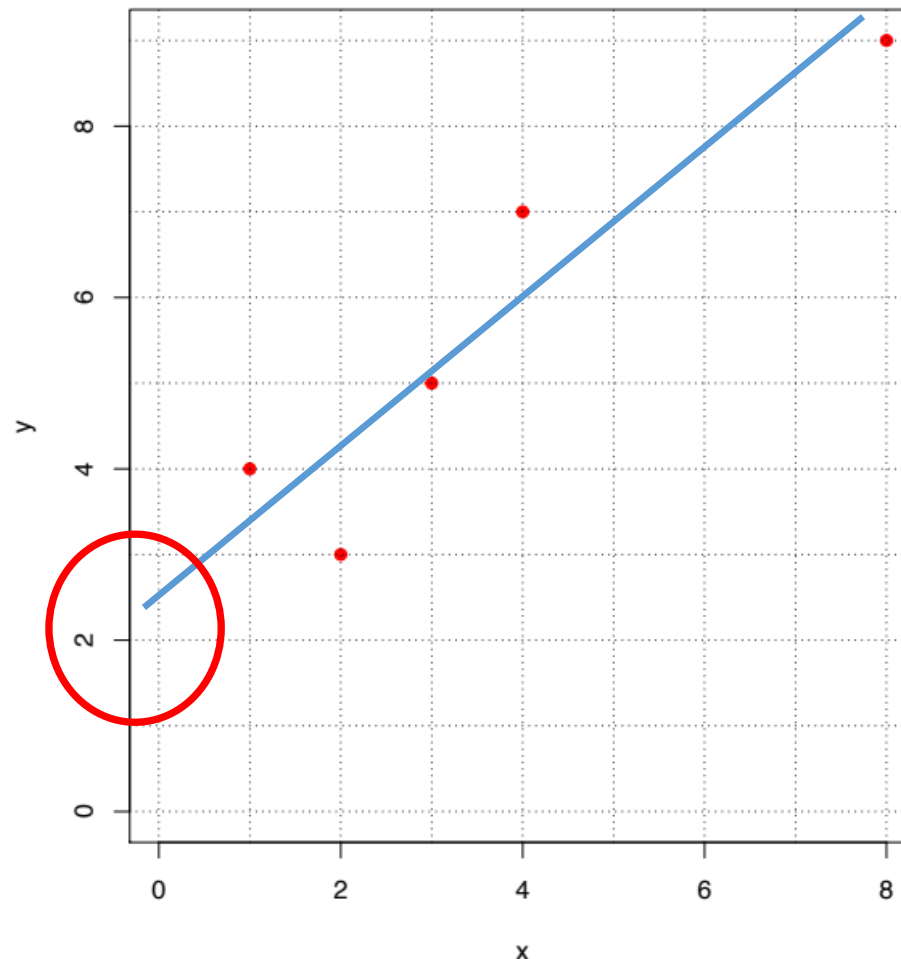
- Let's plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1 :
- Intercept:



$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Let's give this a try...

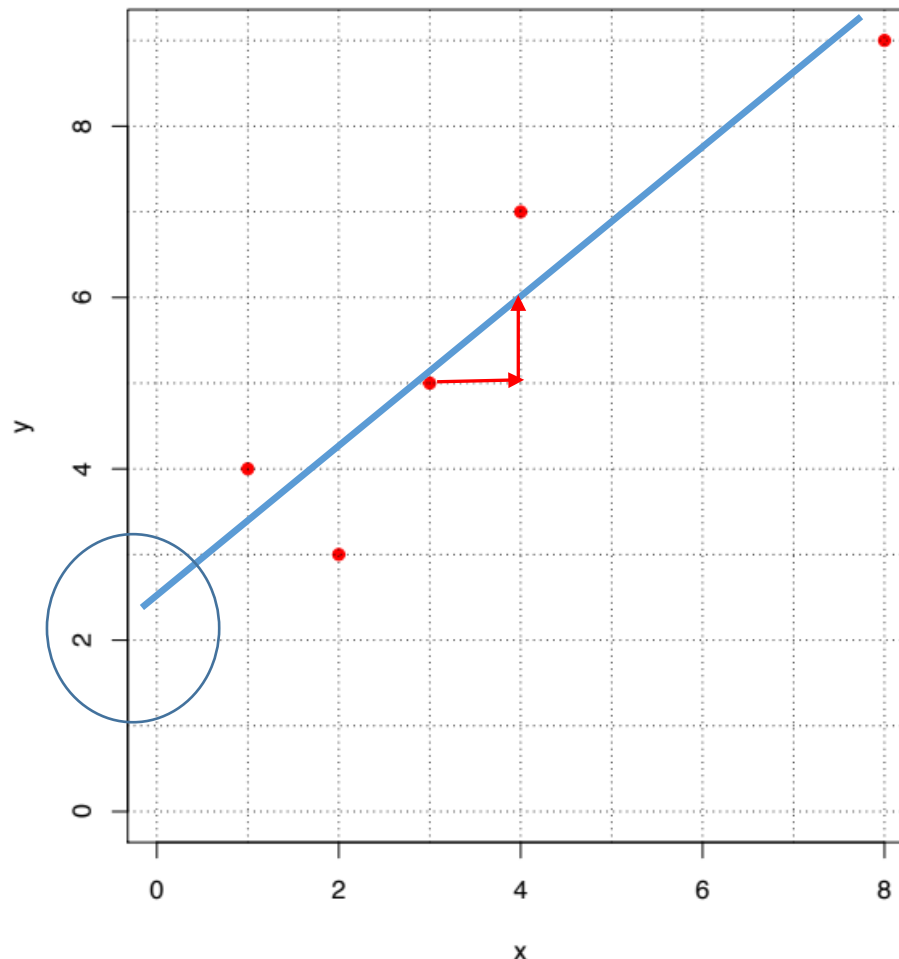
- Let's plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1 :
- Intercept: something 2.2



$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Let's give this a try...

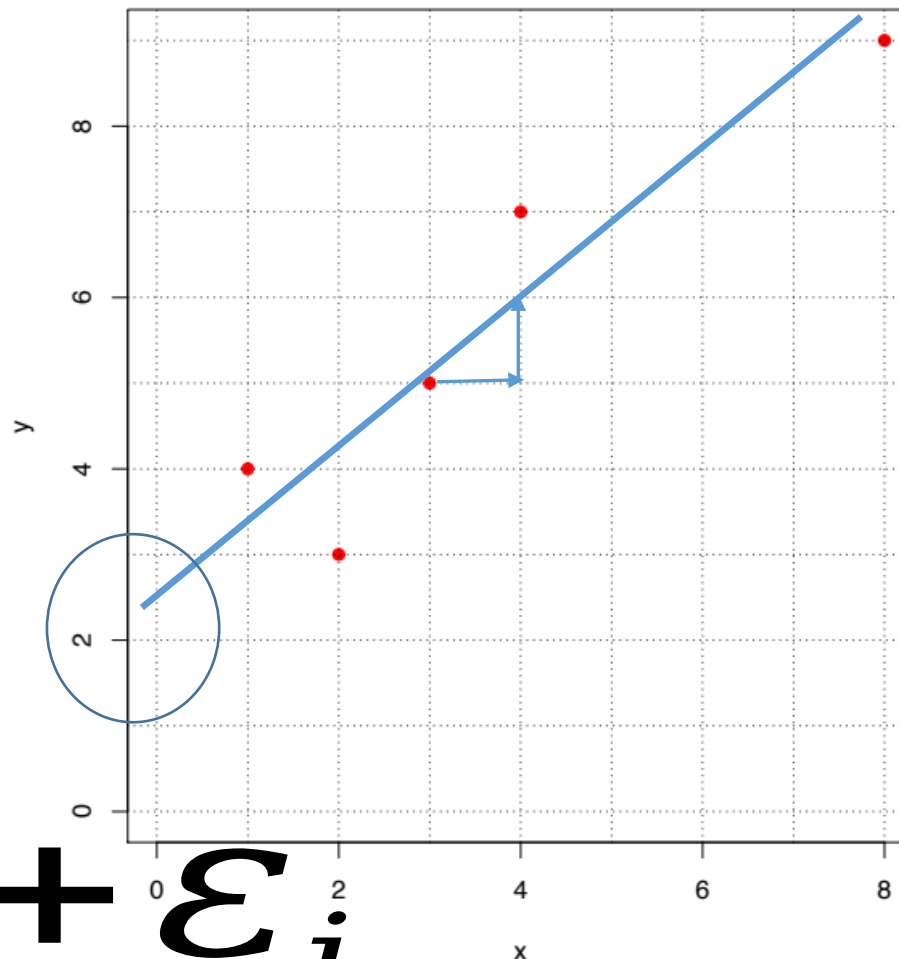
- Let's plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1 :
- Intercept: something 2.2
- Slope: close enough to 1



$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Let's give this a try...

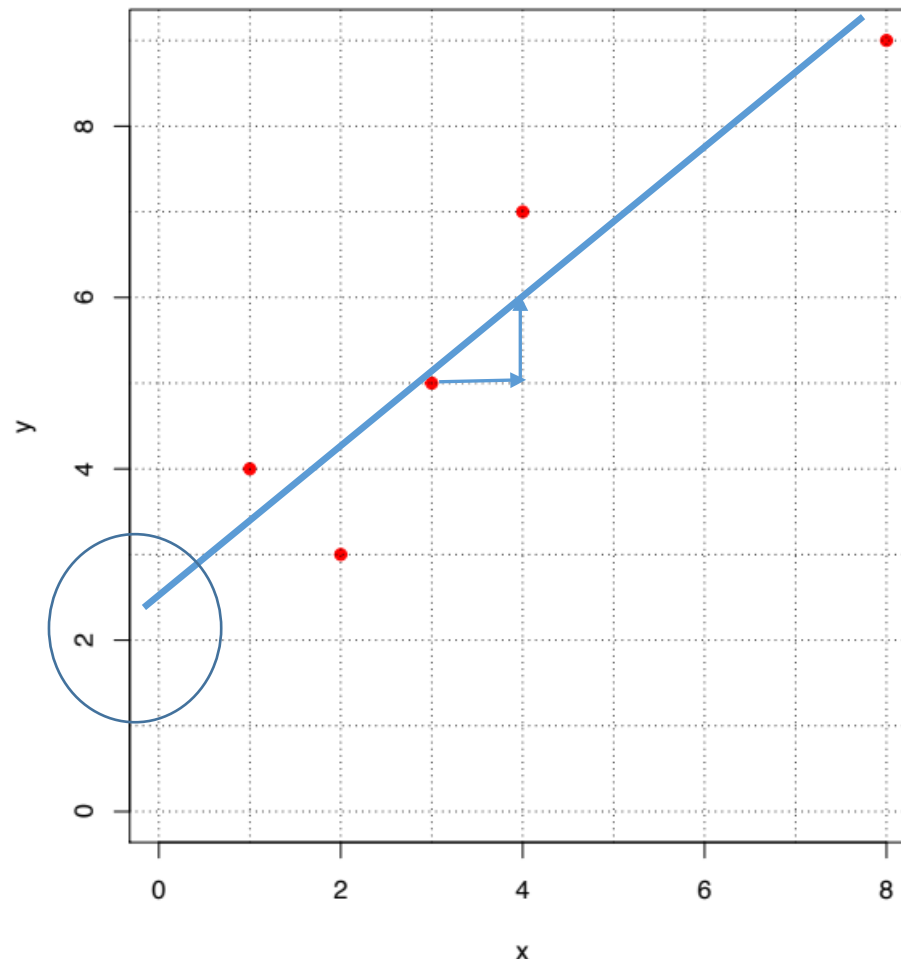
- Let's plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1 :
- Intercept: something 2.2
- Slope: close enough to 1



$$y_i = 2.2 + 1x_i + \varepsilon_i$$
$$y_i = b_0 + b_1x_i + \varepsilon_i$$

Let's give this a try...

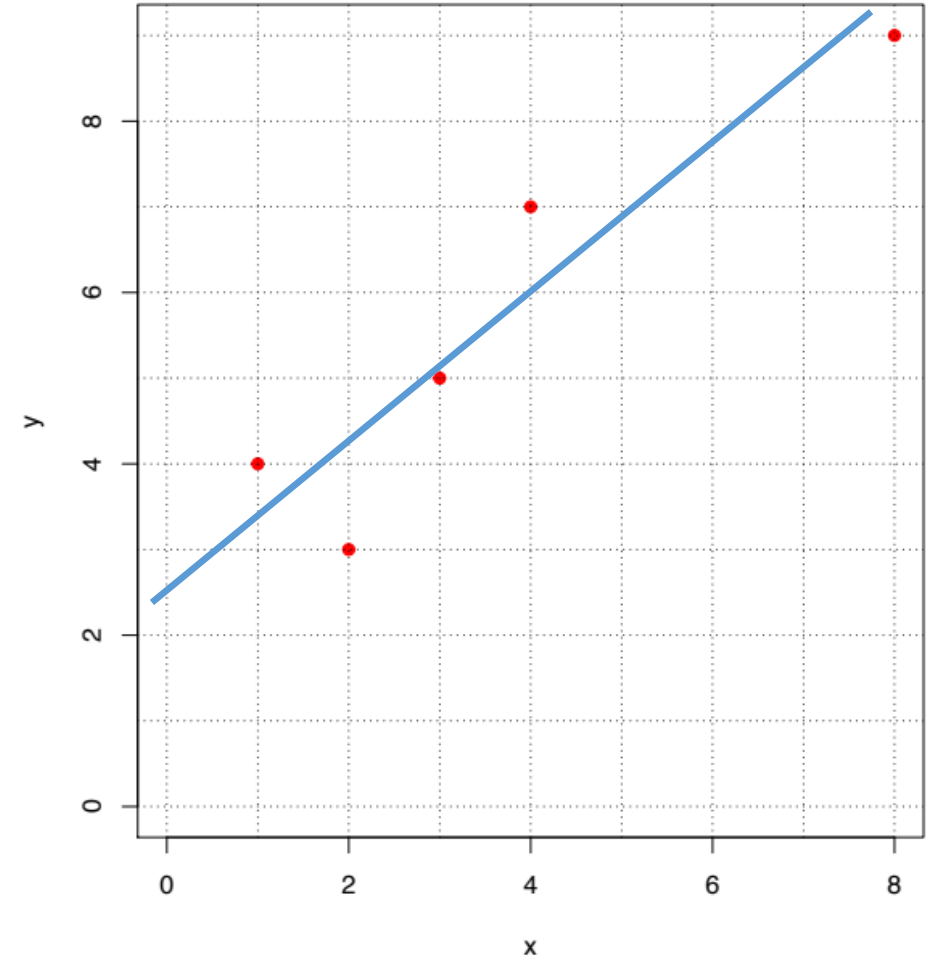
- Let's plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1 :
- Intercept: something 2.2
- Slope: close enough to 1
- But what's with ?



$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

Let's give this a try...

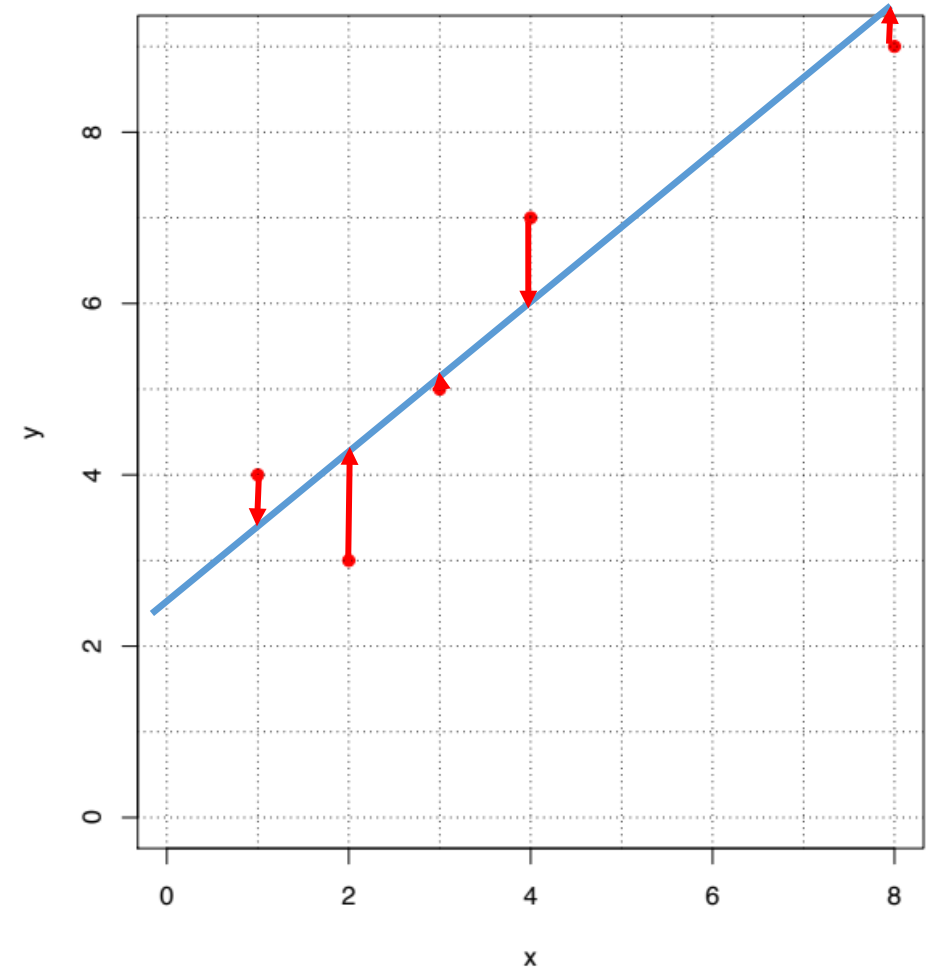
- But what's with ?
- The residuals are the “error” of the model
- We get them by plotting the vertical (y) distance:



$$y_i = 2.2 + 1x_i + \varepsilon_i$$

Let's give this a try...

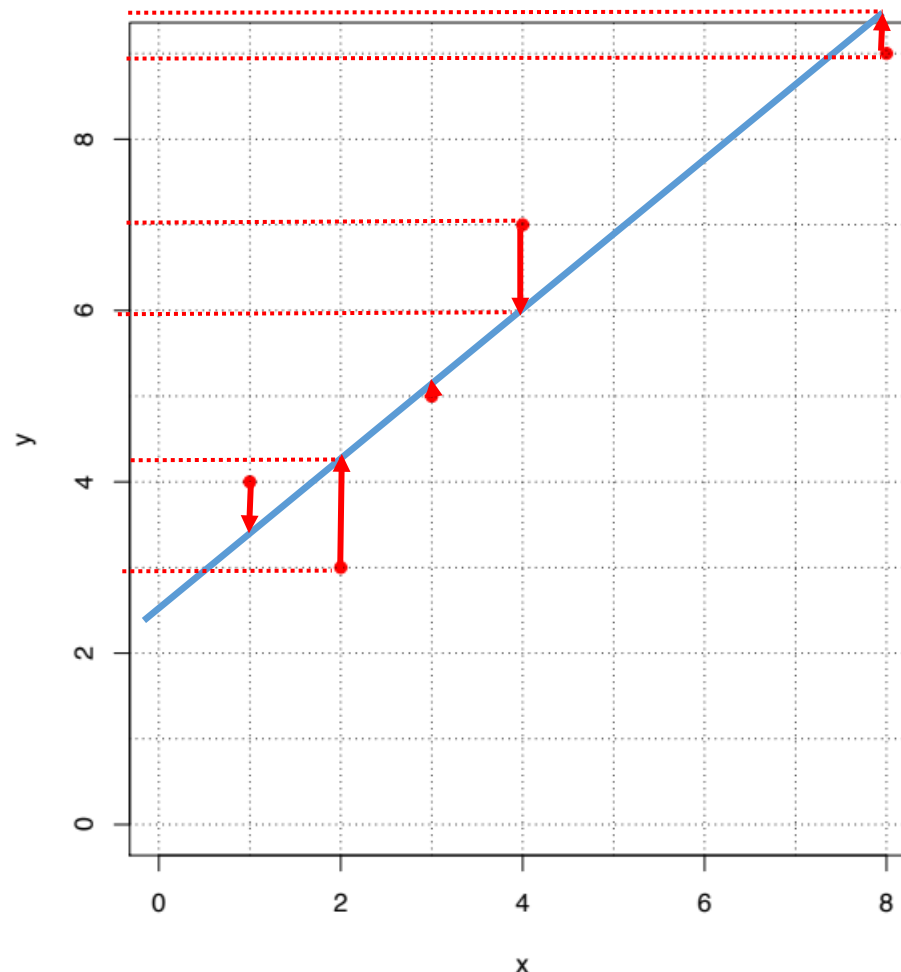
- But what's with ?
- The residuals are the “error” of the model
- We get them by plotting the vertical (y) distance:



$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

Let's give this a try...

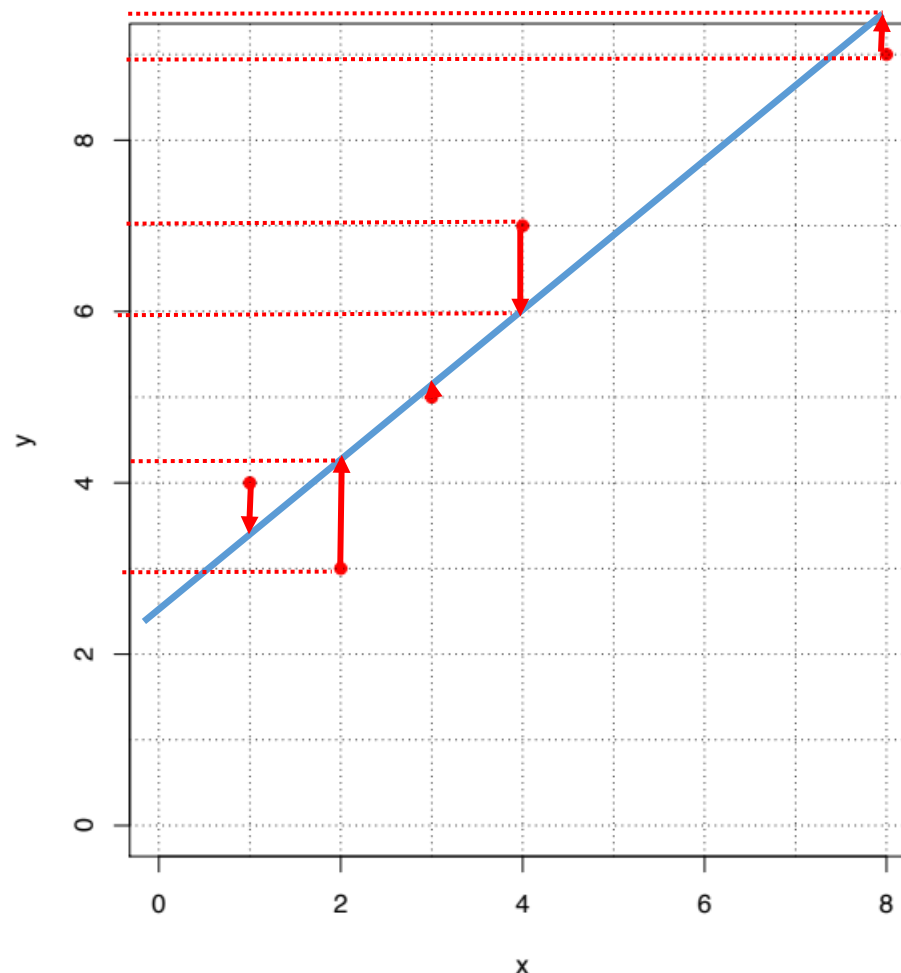
- But what's with ?
- The residuals are the “error” of the model
- We get them by plotting the vertical (y) distance



$$y_i = 2.2 + 1x_i + \epsilon_i$$

Let's give this a try...

- But what's with ?
- The residuals are the “error” of the model
- We get them by plotting the vertical (y) distance
- Just that R does this all for us
- But, HOW?

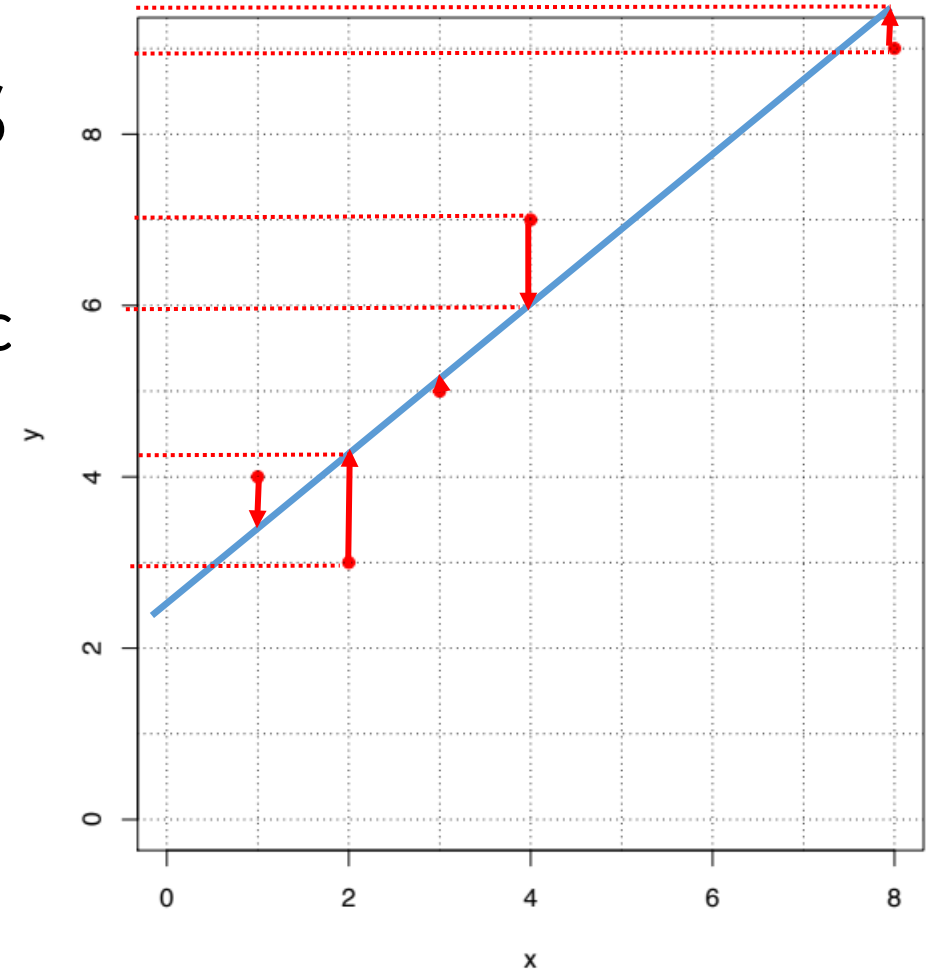


$$y_i = 2.2 + 1x_i + \varepsilon_i$$

-0.7
1.2
0.1
1
0.3

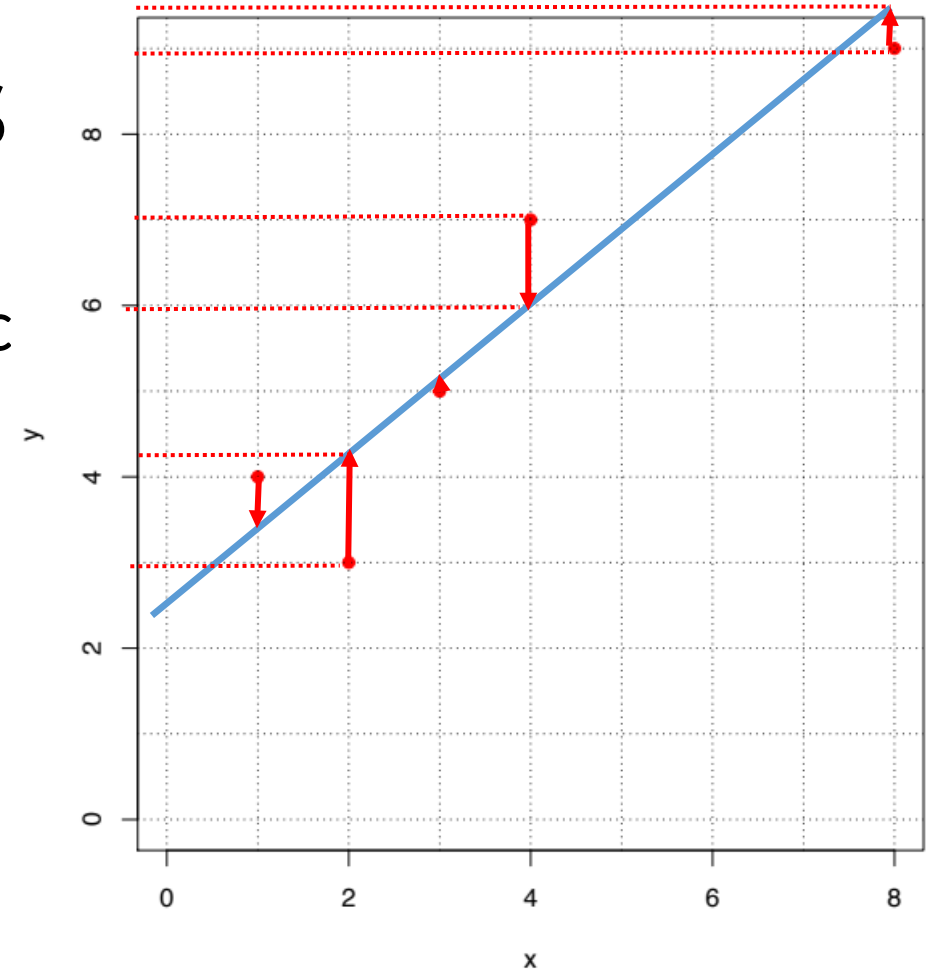
No more guesstimates

- How can we make this process scientific and mathematically tractable?

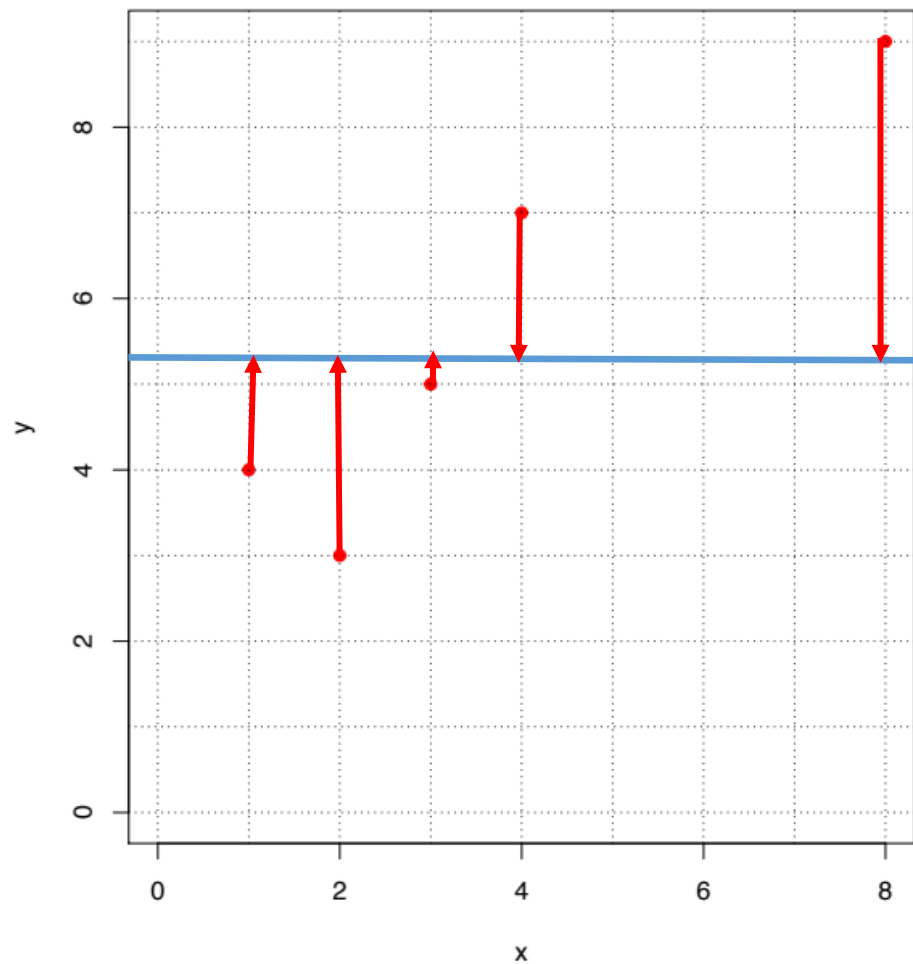


No more guesstimates

- How can we make this process scientific and mathematically tractable?
- Idea: line with smallest residuals wins!

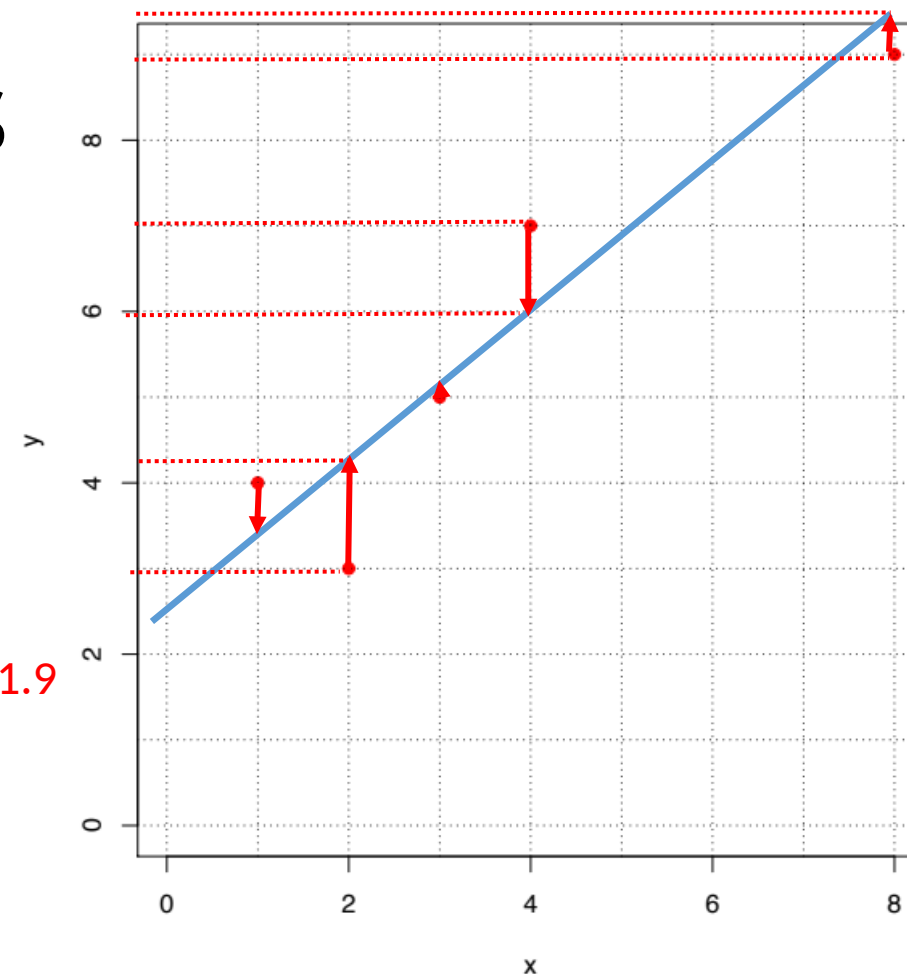


No more quesstimates

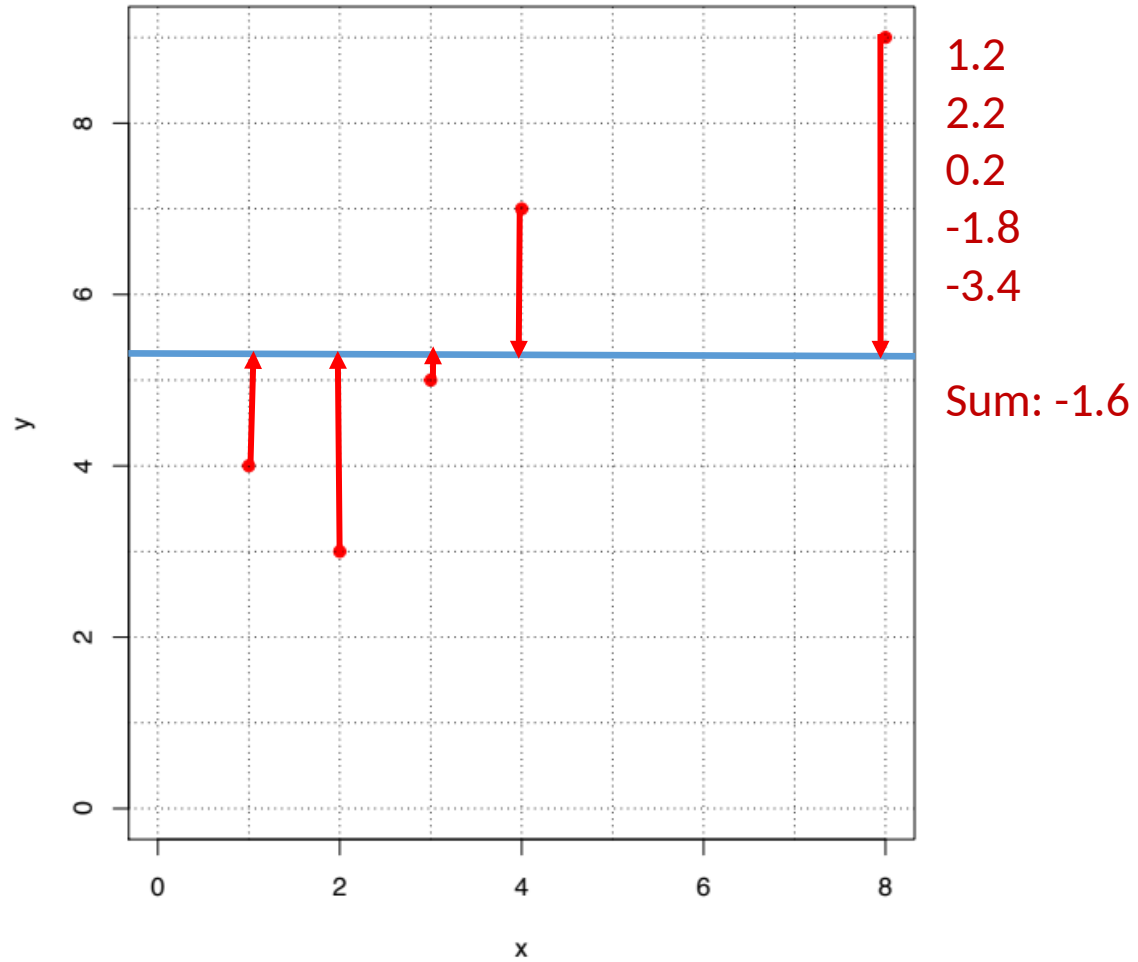


-0.7
1.2
0.1
1
0.3

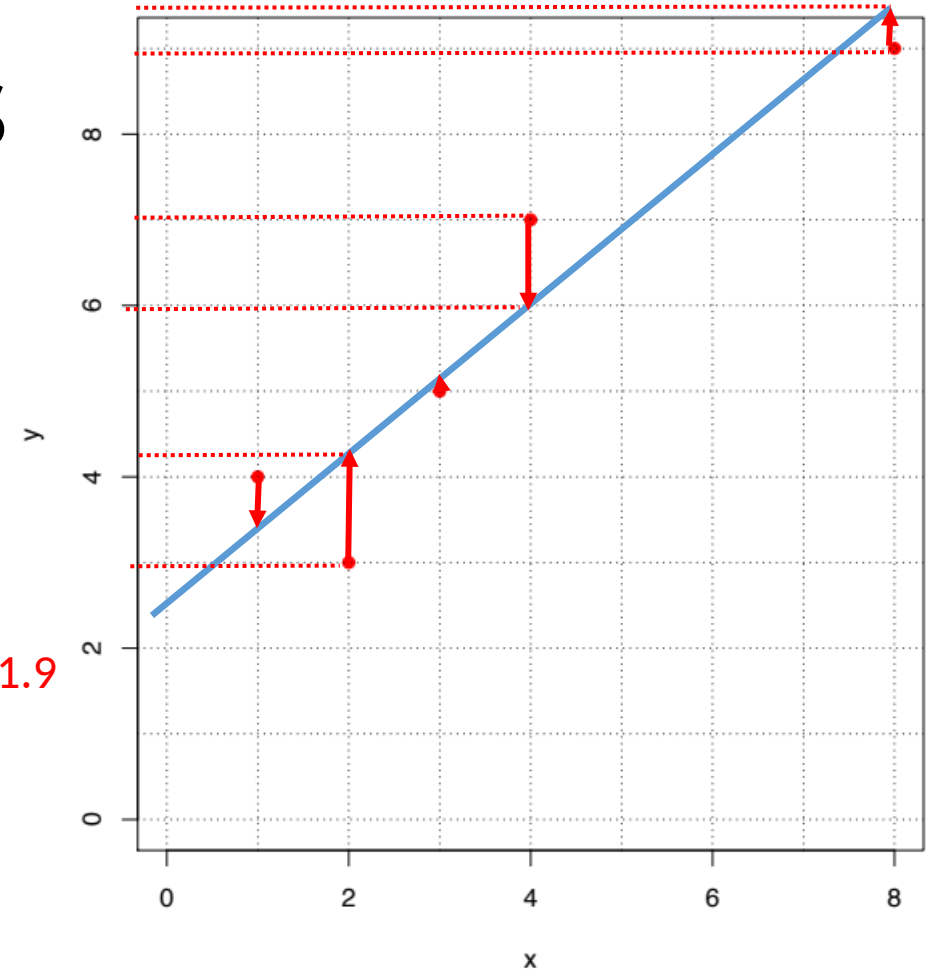
Sum: 1.9



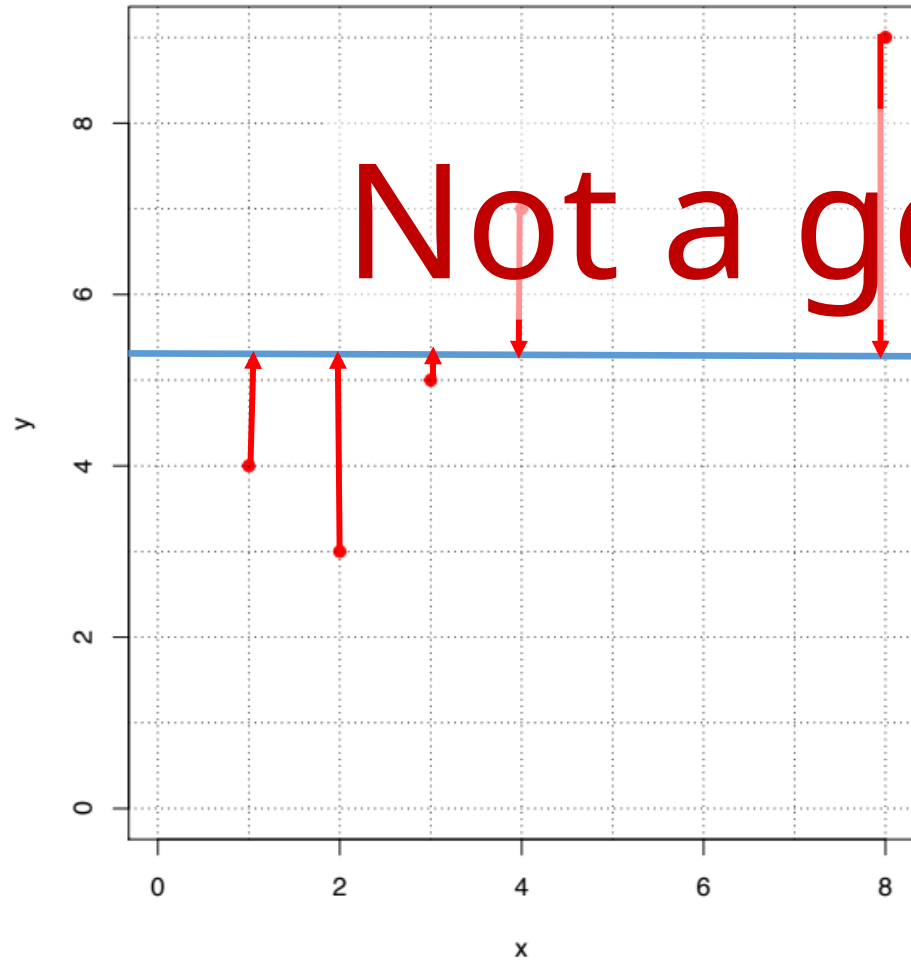
No more quesstimates



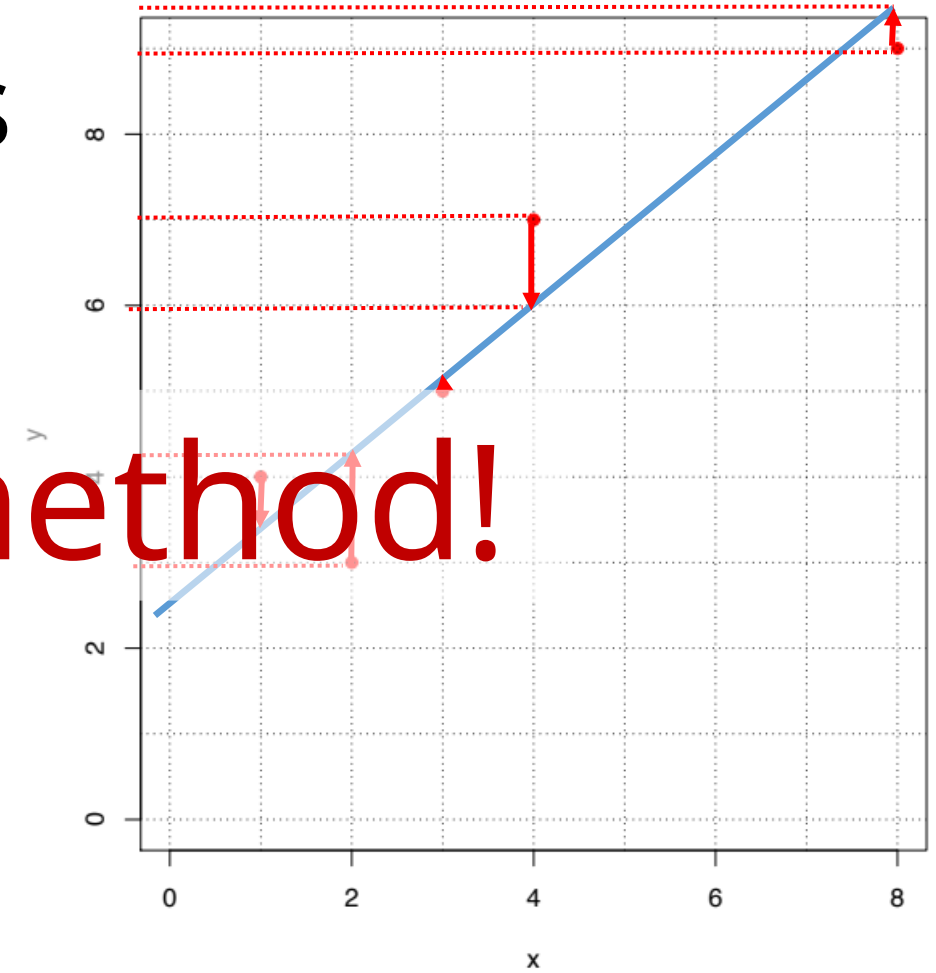
-0.7
1.2
0.1
1
0.3
Sum: 1.9



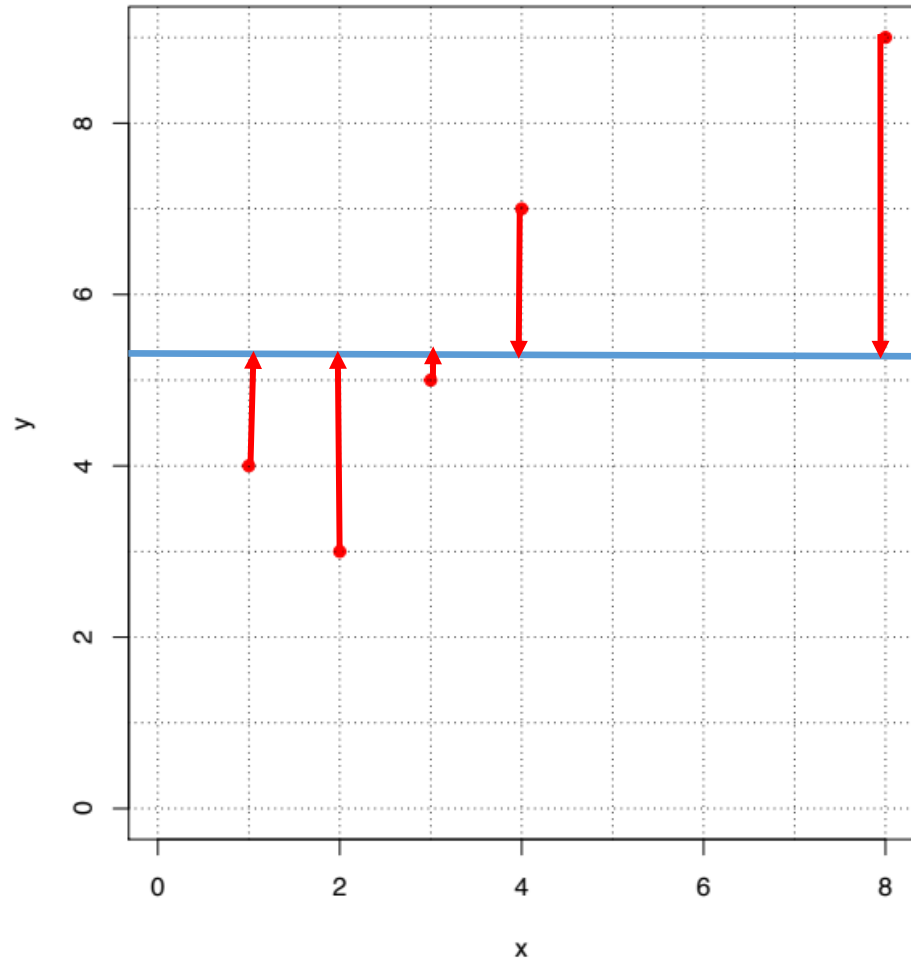
No more quesstimates



Not a good method!

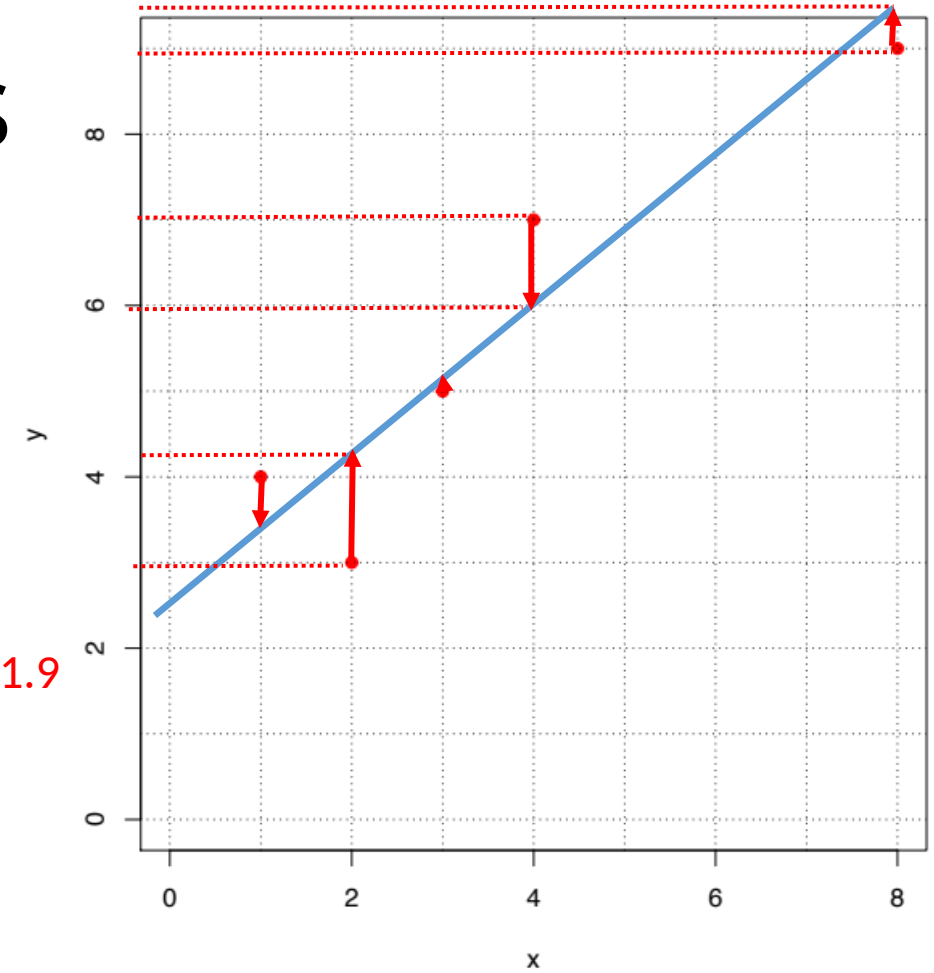


No more quesstimates



1.2
2.2
0.2
-1.8
-3.4
Sum: -1.6

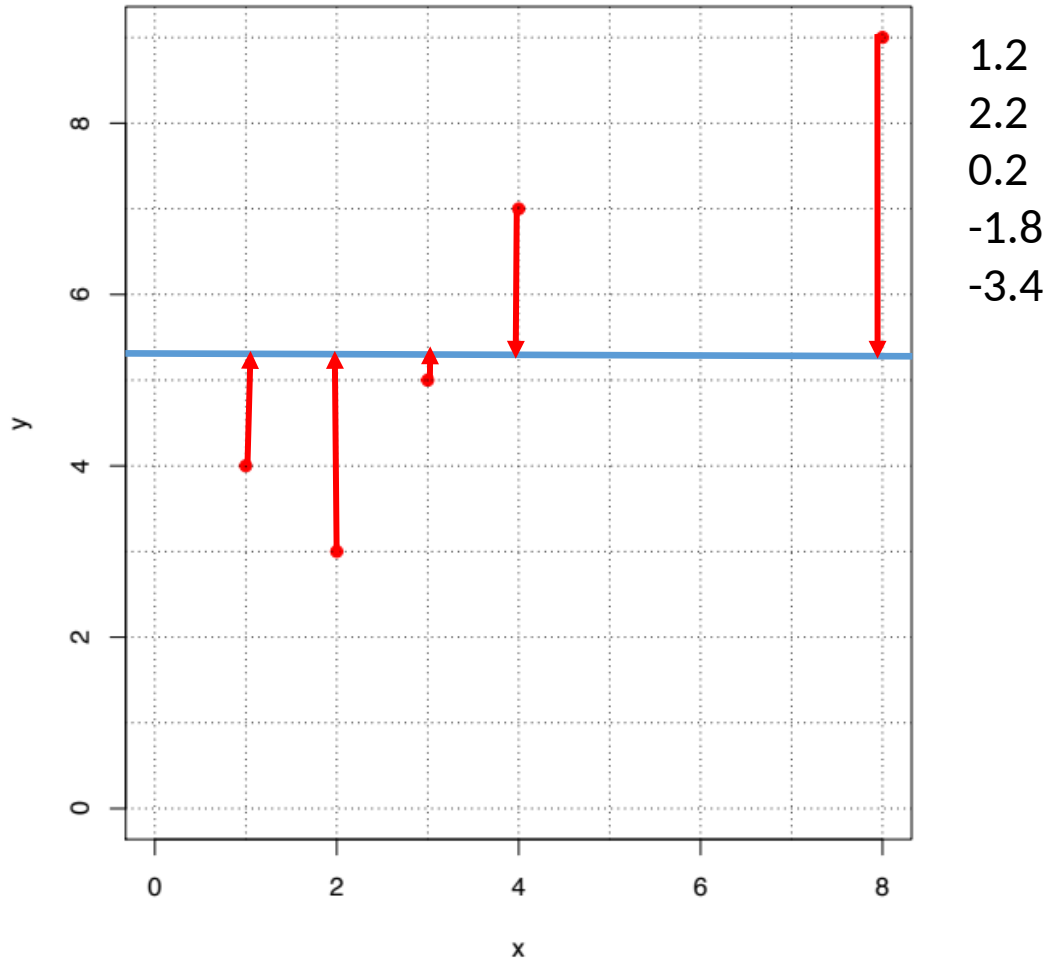
-0.7
1.2
0.1
1
0.3
Sum: 1.9



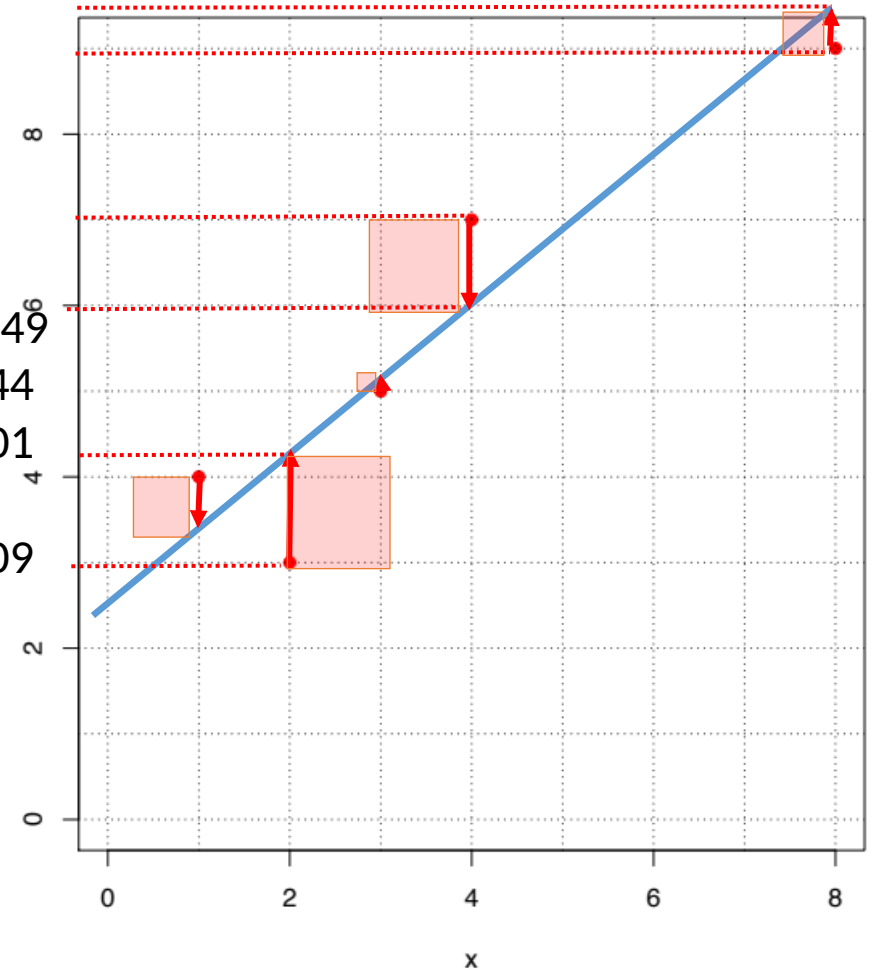
Need to ensure residuals are evaluated as absolute value

Square all of the residuals!

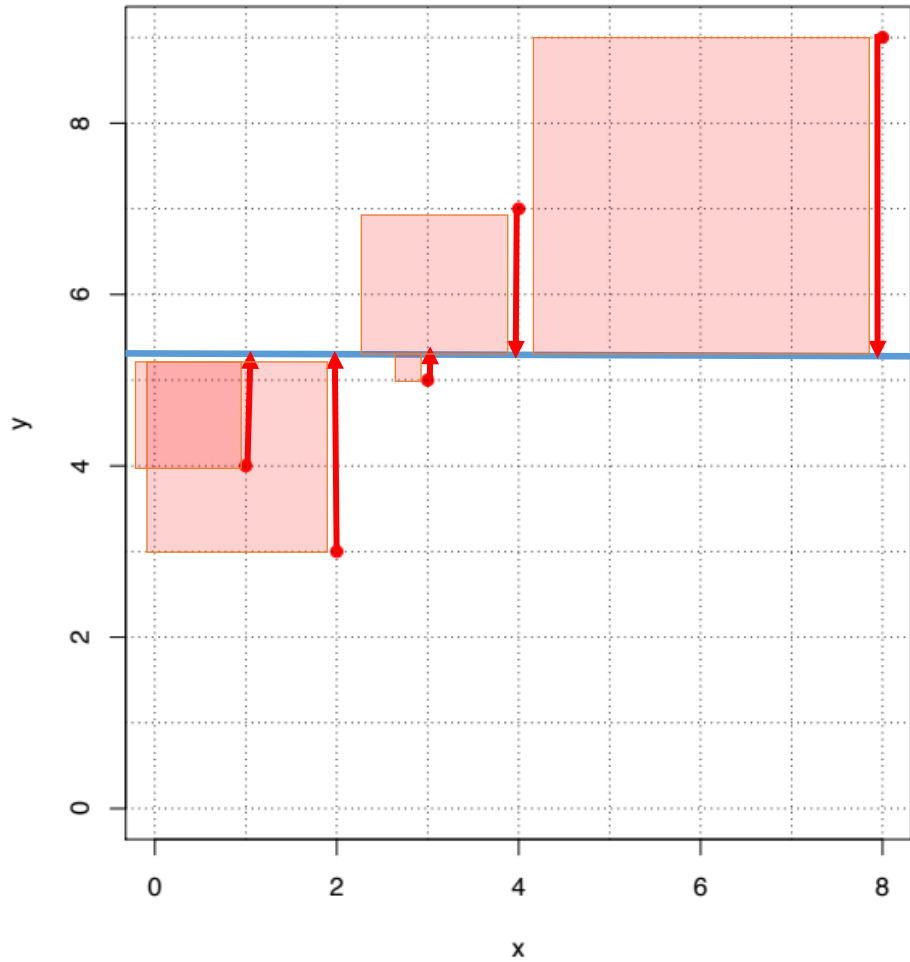
Sums of squares:



$-0.7 \rightarrow 0.49$
 $1.2 \rightarrow 1.44$
 $0.1 \rightarrow 0.01$
 $1 \rightarrow 1$
 $0.3 \rightarrow 0.09$



Sums of squares:



1.2 -> 1.44

2.2 -> 4.84

0.2 -> 0.04

-1.8 -> 3.24

-3.4 -> 11.56

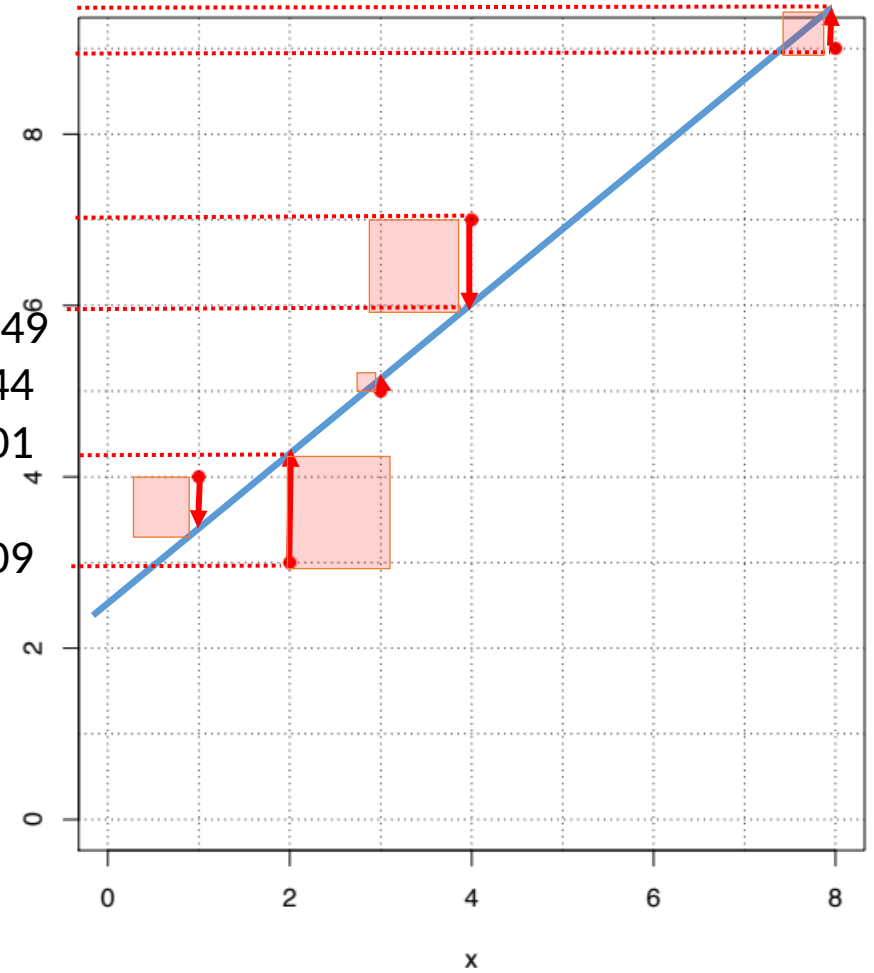
-0.7 -> 0.49

1.2 -> 1.44

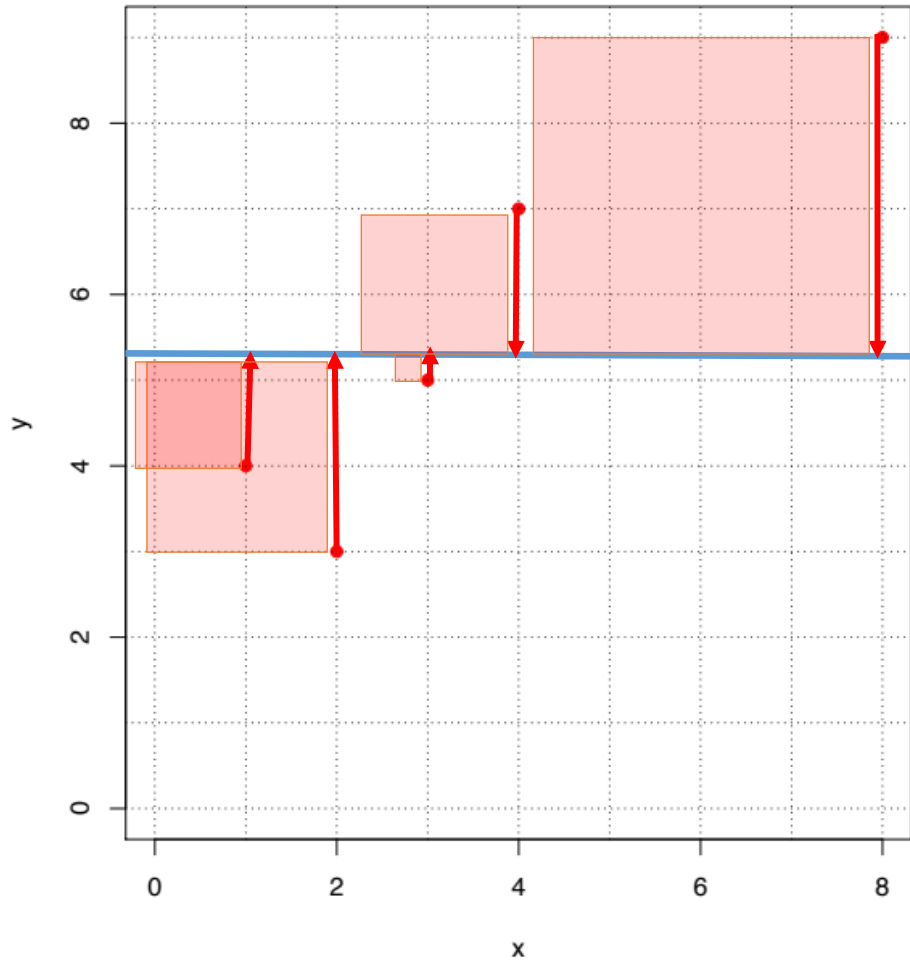
0.1 -> 0.01

1 -> 1

0.3 -> 0.09



Sums of squares:



1.2 \rightarrow 1.44

2.2 \rightarrow 4.84

0.2 \rightarrow 0.04

-1.8 \rightarrow 3.24

-3.4 \rightarrow 11.56

SS: 21.12

-0.7 \rightarrow 0.49

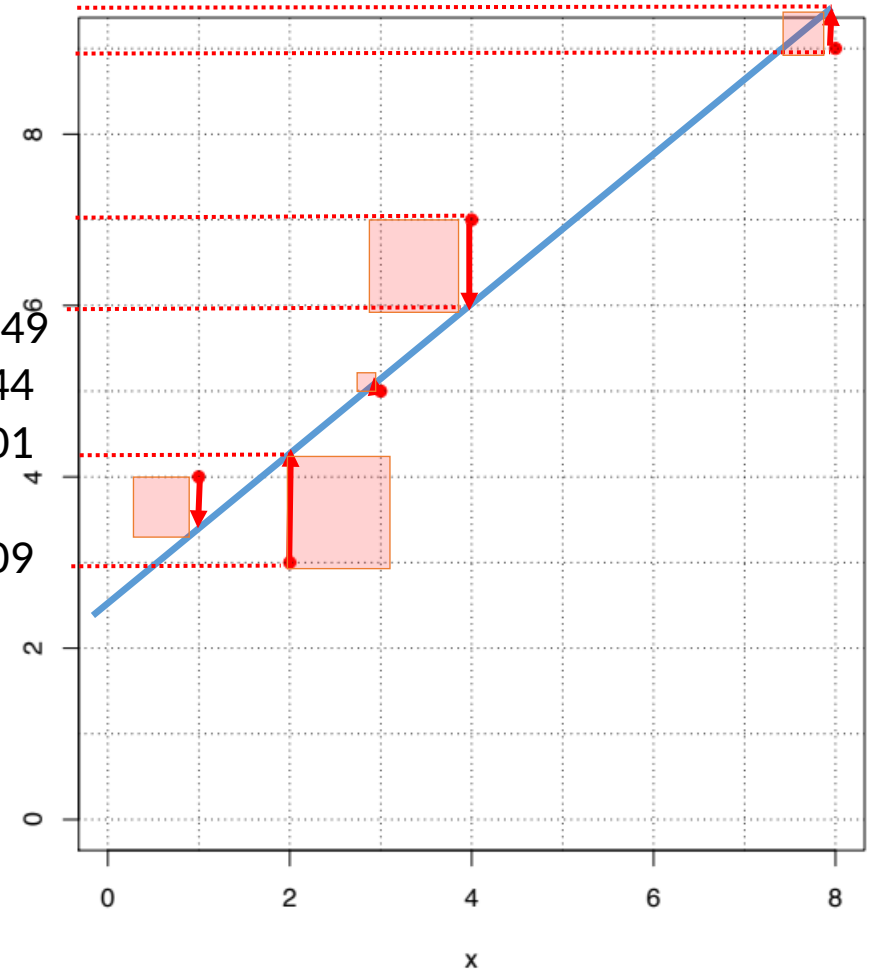
1.2 \rightarrow 1.44

0.1 \rightarrow 0.01

1 \rightarrow 1

0.3 \rightarrow 0.09

SS: 3.03



Use Sums of Squares to assess goodness of fit

- Find b_0 and b_1 of a line that is positioned so that it minimizes the sum of the squared residuals for the data y_i and x_i

Use Sums of Squares to assess goodness of fit

- Find b_0 and b_1 of a line that is positioned so that it minimizes the sum of the squared residuals for the data y_i and x_i
- Solve for b_1 , and b_0

Use Sums of Squares to assess goodness of fit

- Find b_0 and b_1 of a line that is positioned so that it minimizes the sum of the squared residuals for the data y_i and x_i
- Solve for b_1 , and b_0

Use Sums of Squares to assess goodness of fit

- Find b_0 and b_1 of a line that is positioned so that it minimizes the sum of the squared residuals for the data y_i and x_i

ALGEBRA

- Solve for b_1 , and b_0

Use Sums of Squares to assess goodness of fit

- Find b_0 and b_1 of a line that is positioned so that it minimizes the sum of the squared residuals for the data y_i and x_i

ALGEBRA

- Solve for b_1 , and b_0

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

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

Use Sums of Squares to assess goodness of fit

- Find b_0 and b_1 of a line that is positioned so that it minimizes the sum of the squared residuals for the data y_i and x_i

ALGEBRA

- Solve for b_1 , and b_0

$$b_0 = \bar{y} - b_1 \bar{x} \qquad b_1 = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2}$$

Use Sums of Squares to assess goodness of fit

- Find b_0 and b_1 of a line that is positioned so that it minimizes the sum of the squared residuals for the data y_i and x_i

ALGEBRA

- Solve for b_1 , and b_0

$$b_0 = \bar{y} - b_1 \bar{x}$$
$$b_1 = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2}$$
$$= \frac{\text{Cov}[x, y]}{\sigma_x^2}$$

Use Sums of Squares to assess goodness of fit

- Find b_0 and b_1 of a line that is positioned so that it minimizes the sum of the squared residuals for the data y_i and x_i

ALGEBRA

- Solve for b_1 , and b_0

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

$$b_1 = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2}$$

Covariance between x and y

$$i \frac{Cov[x, y]}{\sigma_x^2}$$

Use Sums of Squares to assess goodness of fit

- Find b_0 and b_1 of a line that is positioned so that it minimizes the sum of the squared residuals for the data y_i and x_i

ALGEBRA

- Solve for b_1 , and b_0

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

$$b_1 = \frac{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2}$$

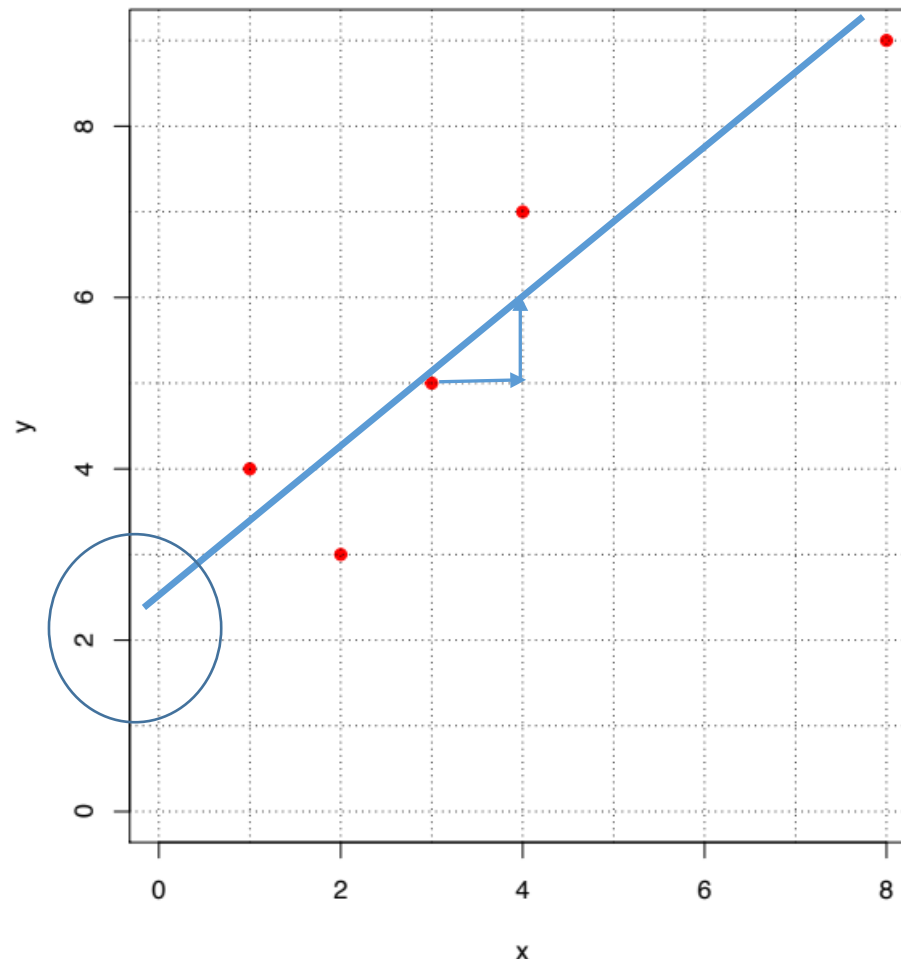
Covariance between x and y

b_1

Variance of x

Let's give this a try...

- Let's plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1 :
- Intercept: something 2.2
- Slope: close enough to 1
- But what's with ?

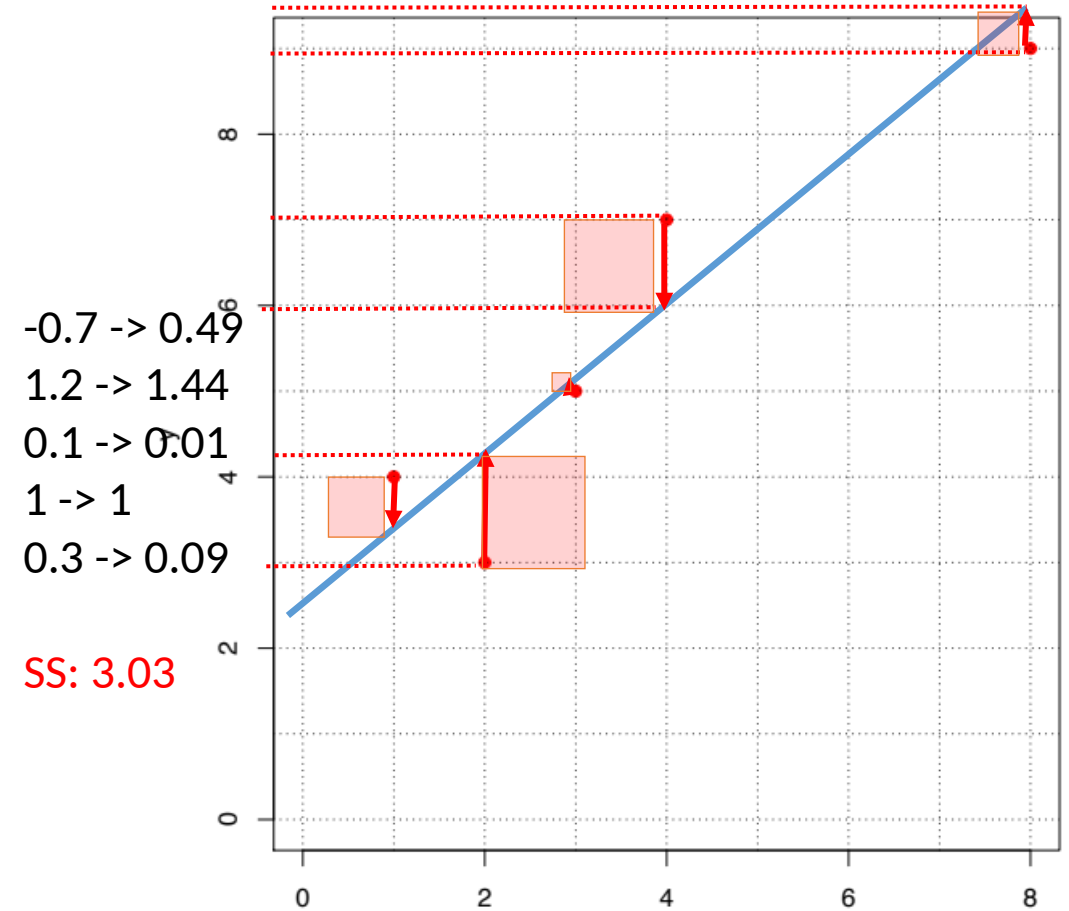


$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

Let's give this a try:

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

b_1



$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

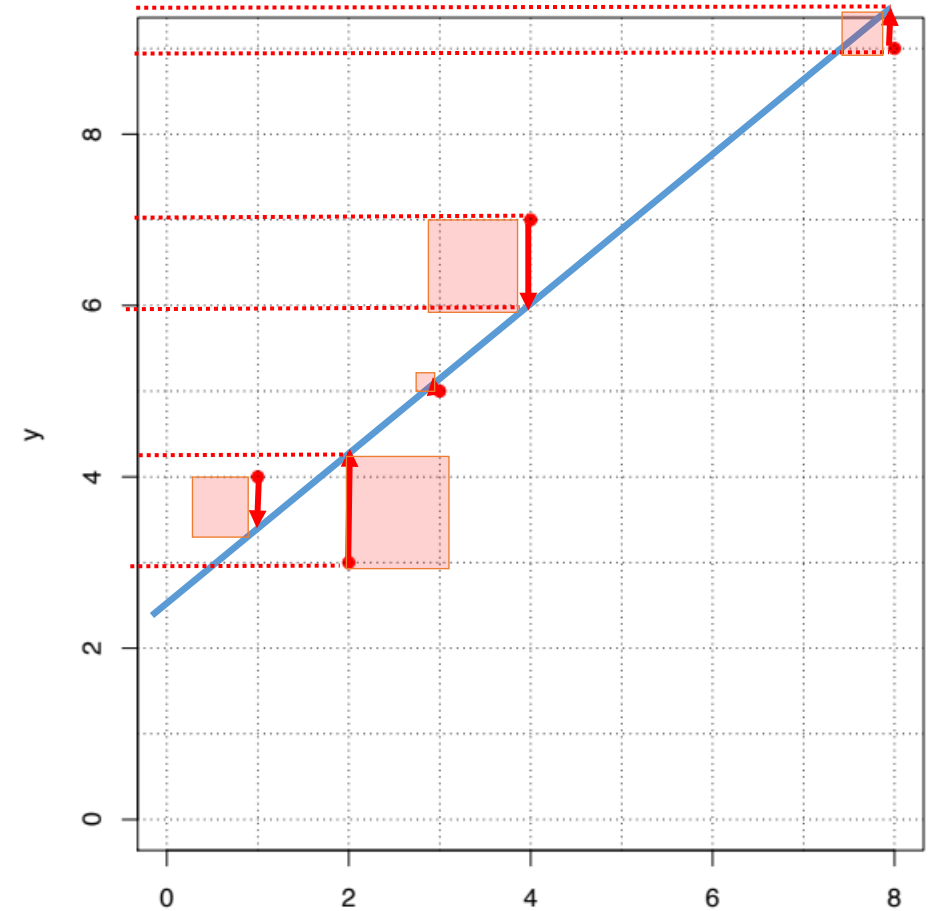
SS: 3.03

Let's give this a try:

x, y
1,4
2,3
3.5
4,7
8,9

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

b_1



$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

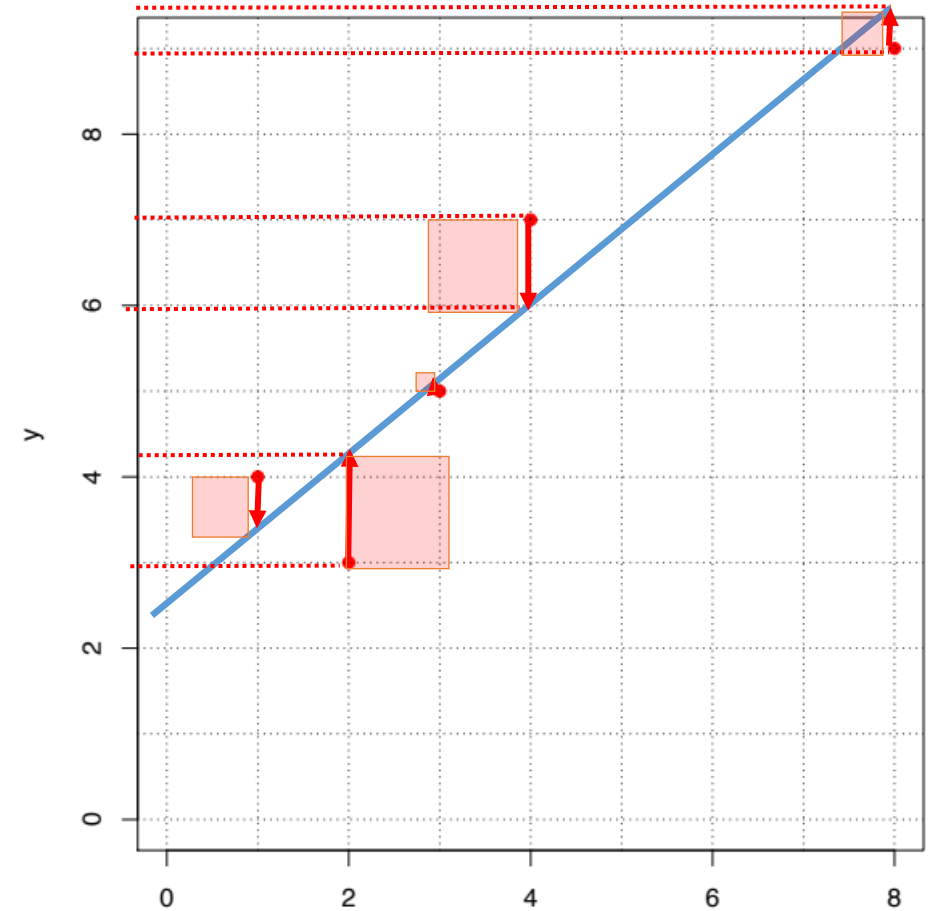
SS: 3.03

Let's give this a try:

x, y
1,4
2,3
3.5
4,7
8,9

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

b_1

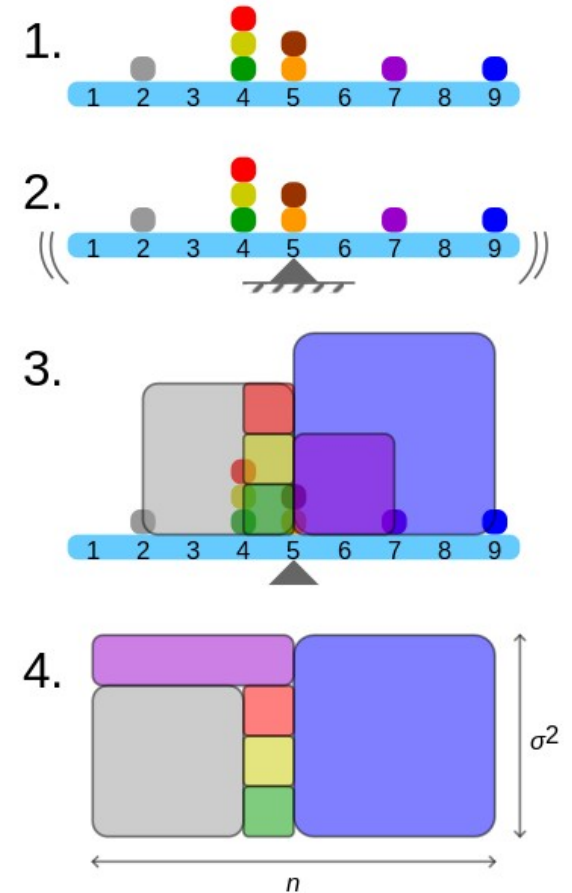
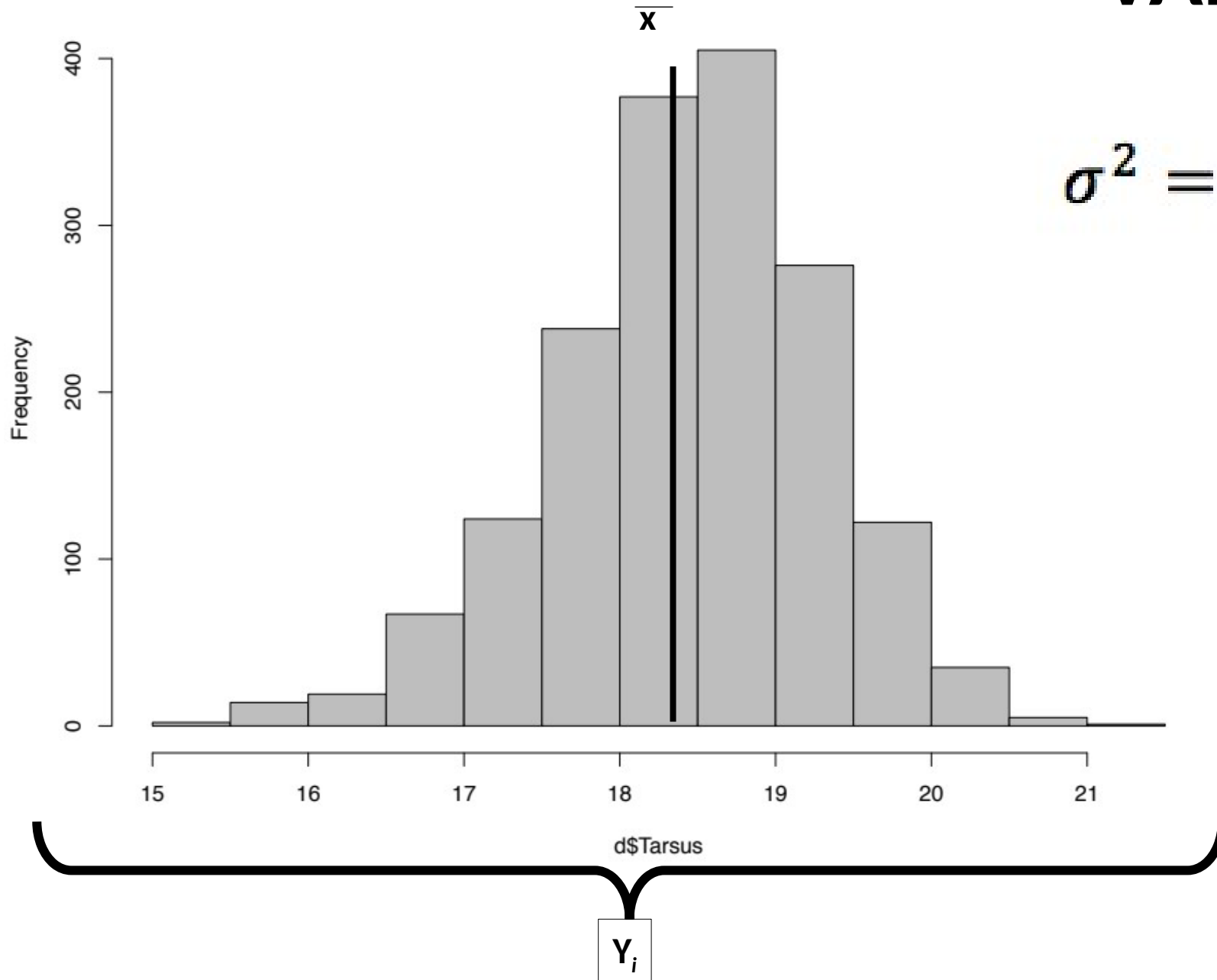


$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

SS: 3.03

VARIANCE

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$



Let's give this a try:

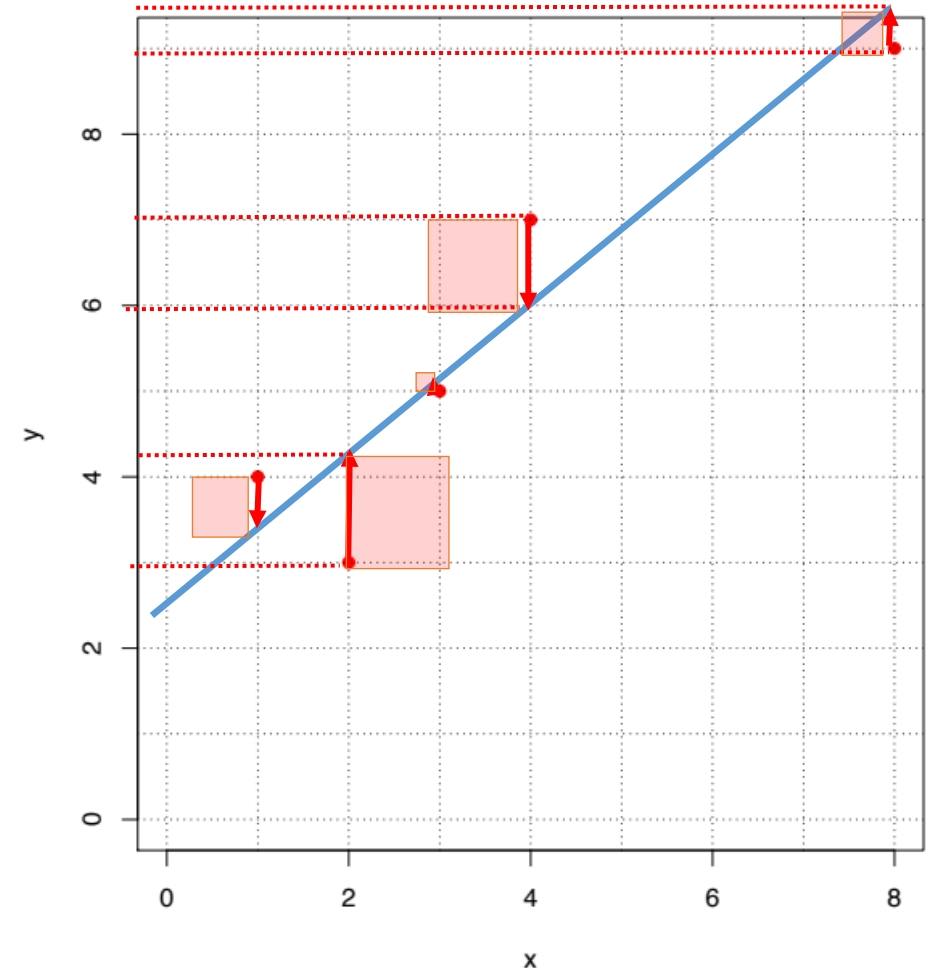
x, y
1,4
2,3
3.5
4,7
8,9

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

=

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

b_1



$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

SS: 3.03

Let's give this a try:

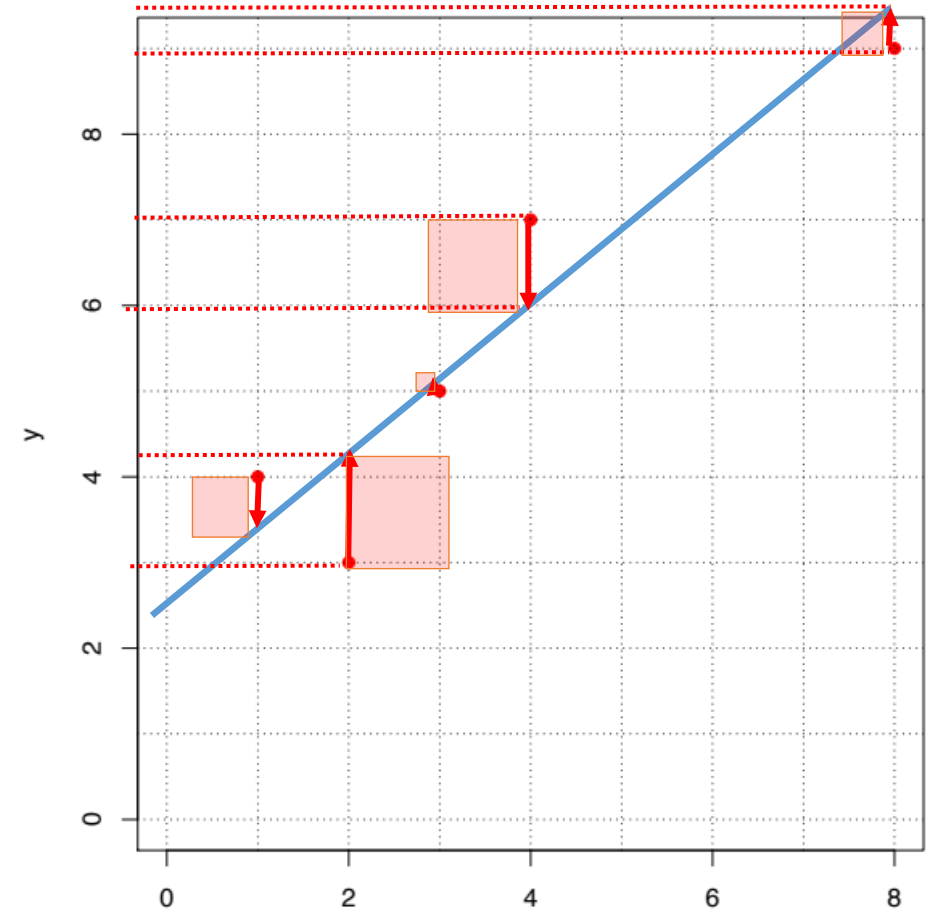
x, y
1,4
2,3
3.5
4,7
8,9

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

=

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

b_1



$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

SS: 3.03

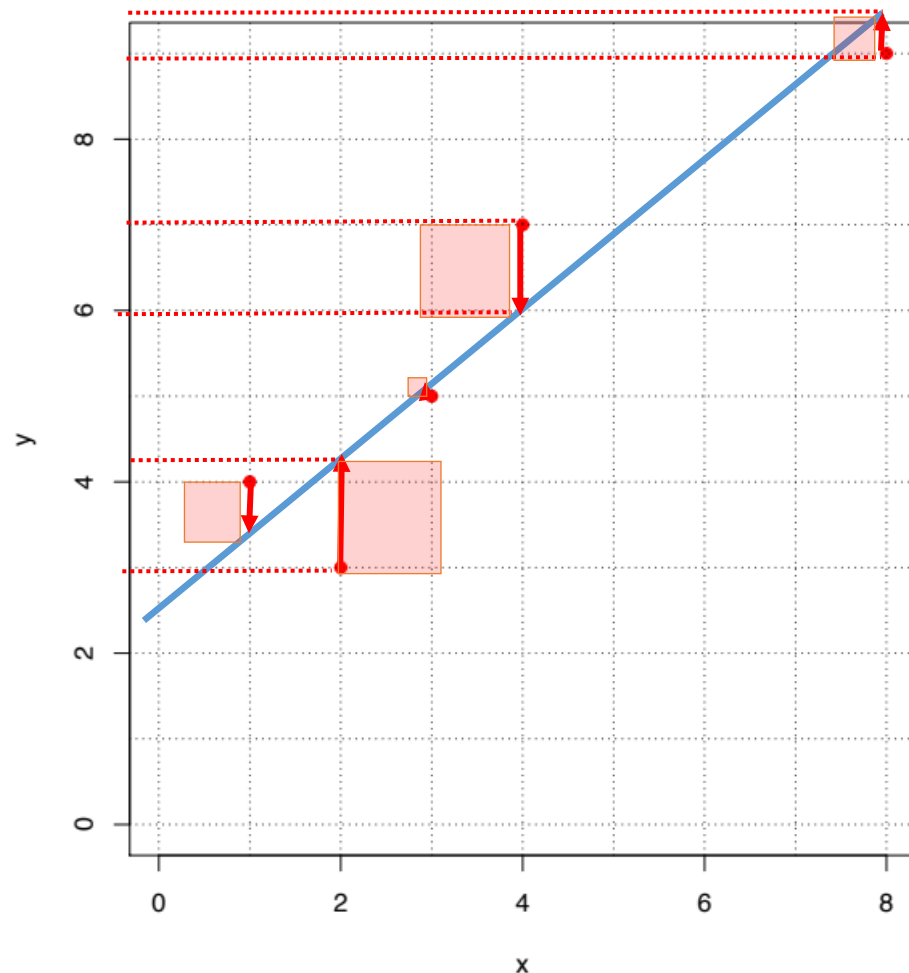
Let's give this a try:

x, y
1,4
2,3
3.5
4,7
8,9

$$\text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

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

b_1



$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

SS: 3.03

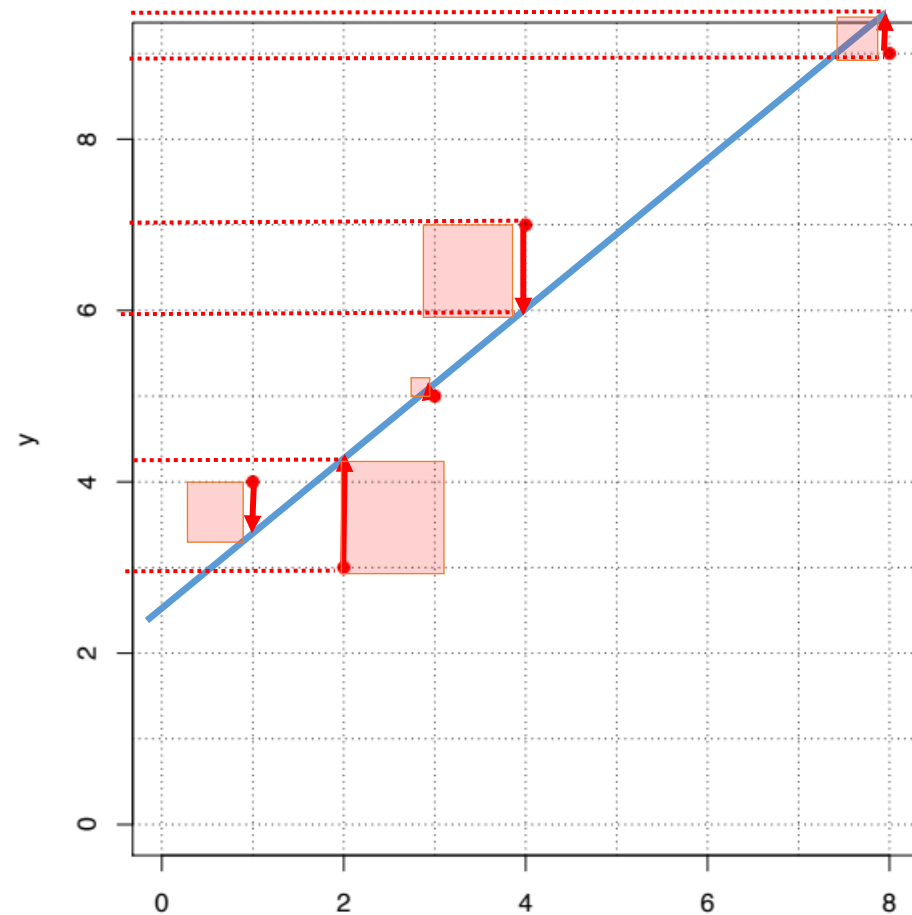
Let's give this a try:

$$\begin{array}{l} x, y \\ 1, 4 \\ 2, 3 \\ 3, 5 \\ 4, 7 \\ 8, 9 \end{array} \quad \text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

=

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

b_1



$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

SS: 3.03

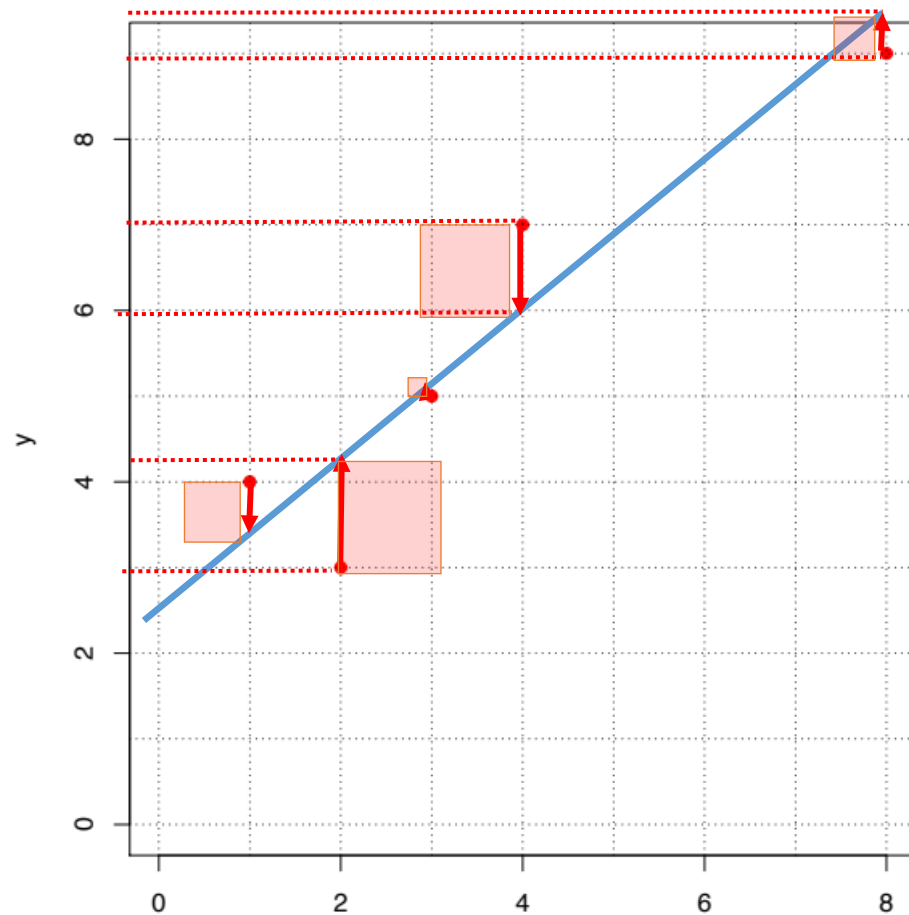
Let's give this a try:

$$\begin{array}{l} x, y \\ 1, 4 \\ 2, 3 \\ 3.5 \\ 4, 7 \\ 8, 9 \end{array} \quad \text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

=

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

b_1



$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

SS: 3.03

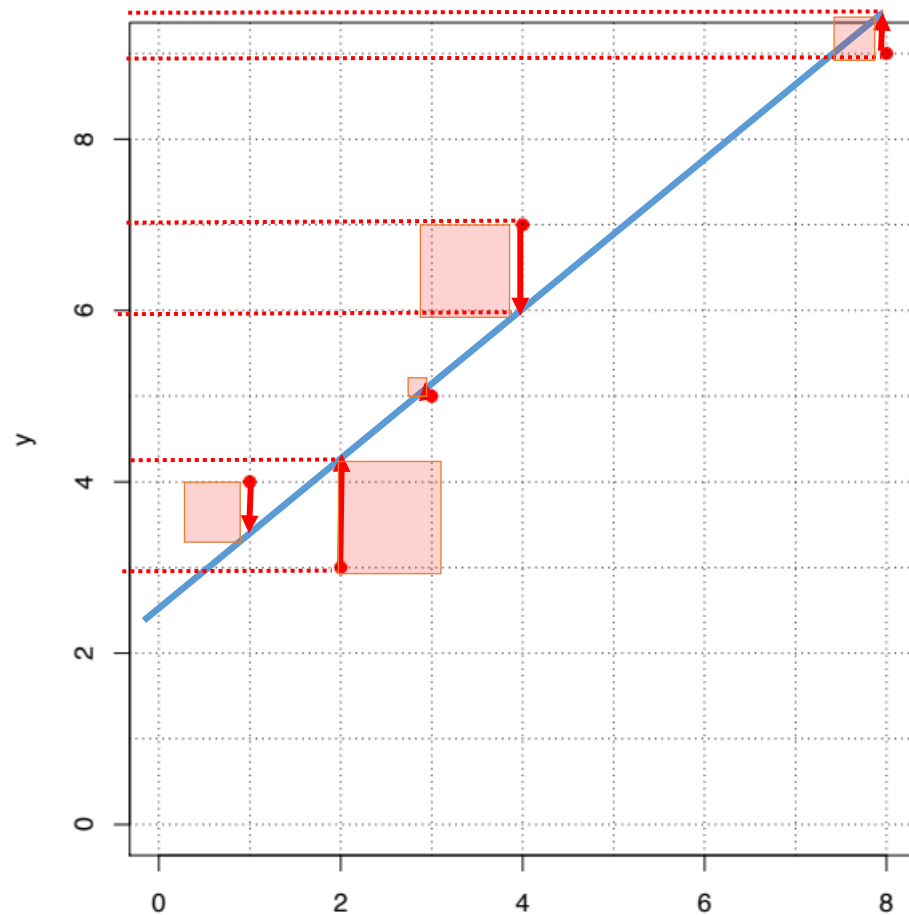
Let's give this a try:

$$\begin{array}{l} x, y \\ 1, 4 \\ 2, 3 \\ 3.5 \\ 4, 7 \\ 8, 9 \end{array} \quad \text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

=

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

b_1



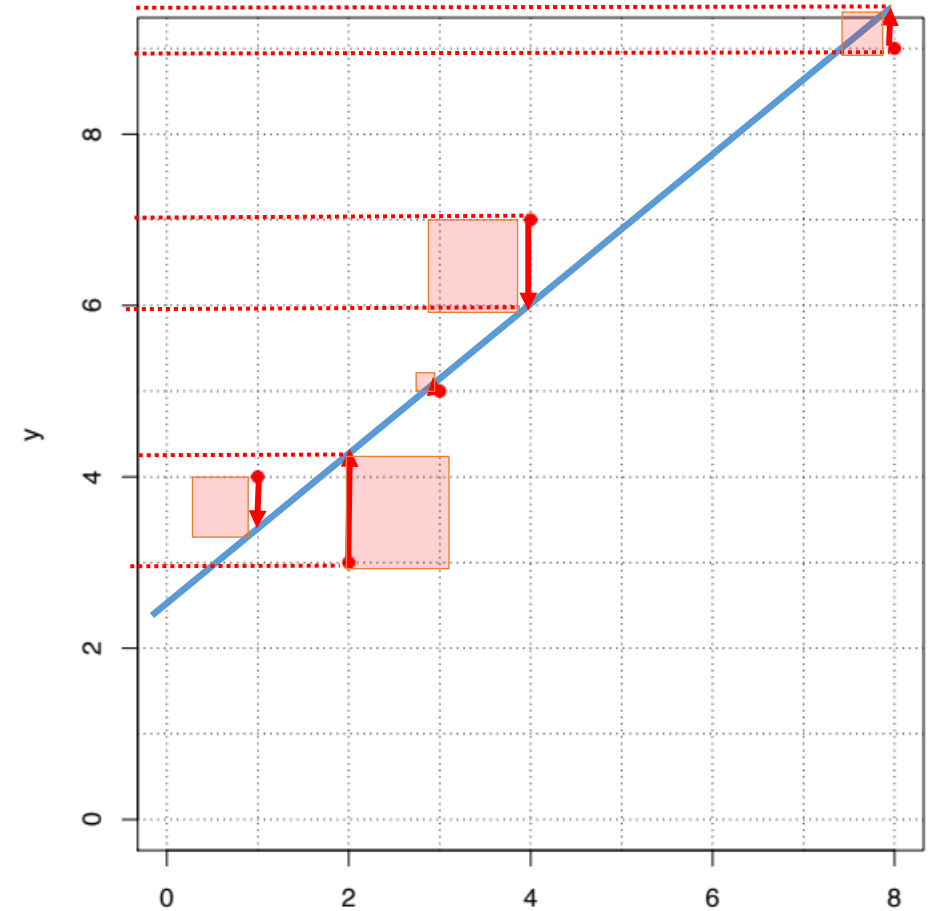
$$y_i = 2.2 + 1 x_i + \varepsilon_i$$

SS: 3.03

Let's give this a try:

$$\begin{array}{l} x, y \\ 1, 4 \\ 2, 3 \\ 3, 5 \\ 4, 7 \\ 8, 9 \end{array} \quad \text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

b_1



$$y_i = 2.2 + 0.83 x_i + \varepsilon_i$$

SS: 3.03

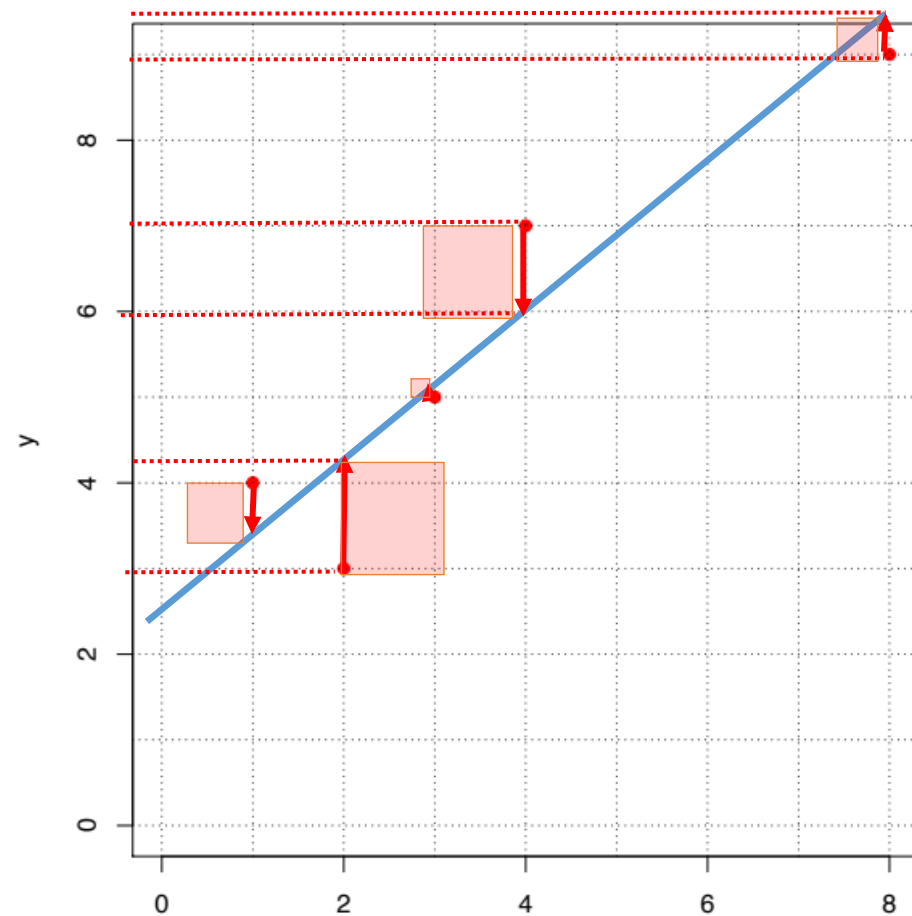
Let's give this a try:

$$\begin{array}{l} x, y \\ 1, 4 \\ 2, 3 \\ 3, 5 \\ 4, 7 \\ 8, 9 \end{array} \quad \text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

=

$$= 5.6 - 0.83 * 3.6$$

b_1



$$y_i = 2.2 + 0.83 x_i + \varepsilon_i$$

SS: 3.03

Let's give this a try:

$$\begin{array}{l} x, y \\ 1, 4 \\ 2, 3 \\ 3, 5 \\ 4, 7 \\ 8, 9 \end{array} \quad \text{Cov}(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

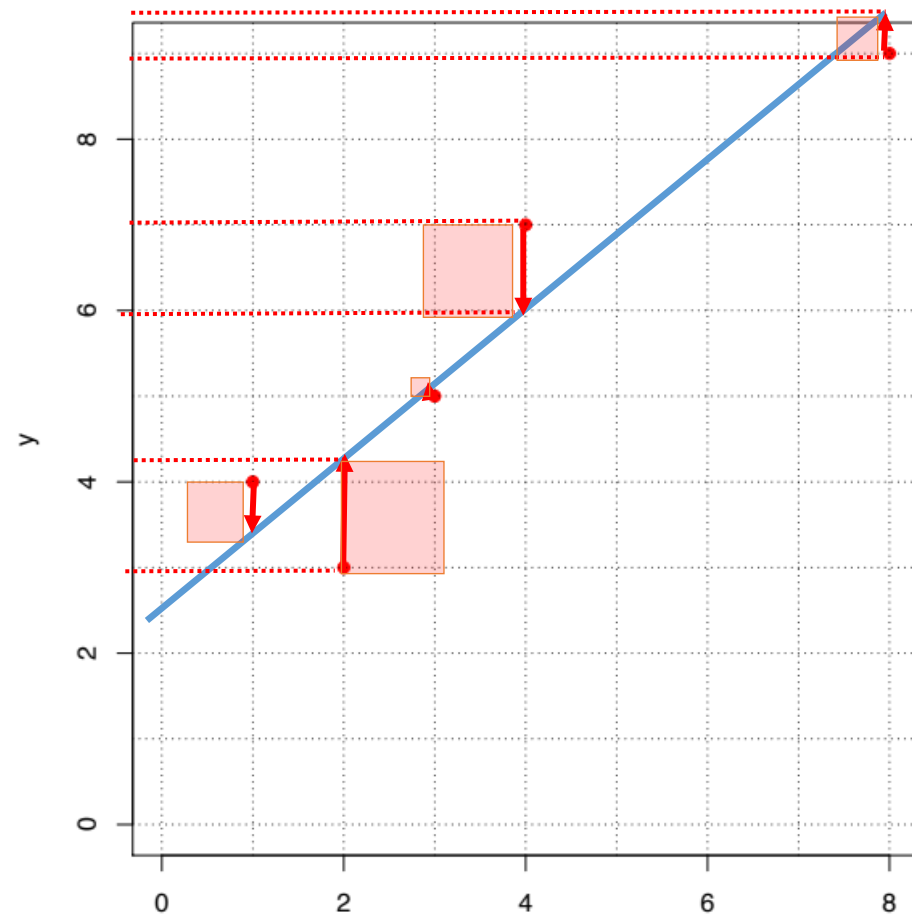
=

$$= 5.6 - 0.83 * 3.6$$

$$= 5.6 - 2.99$$

$$= 2.5$$

b_1



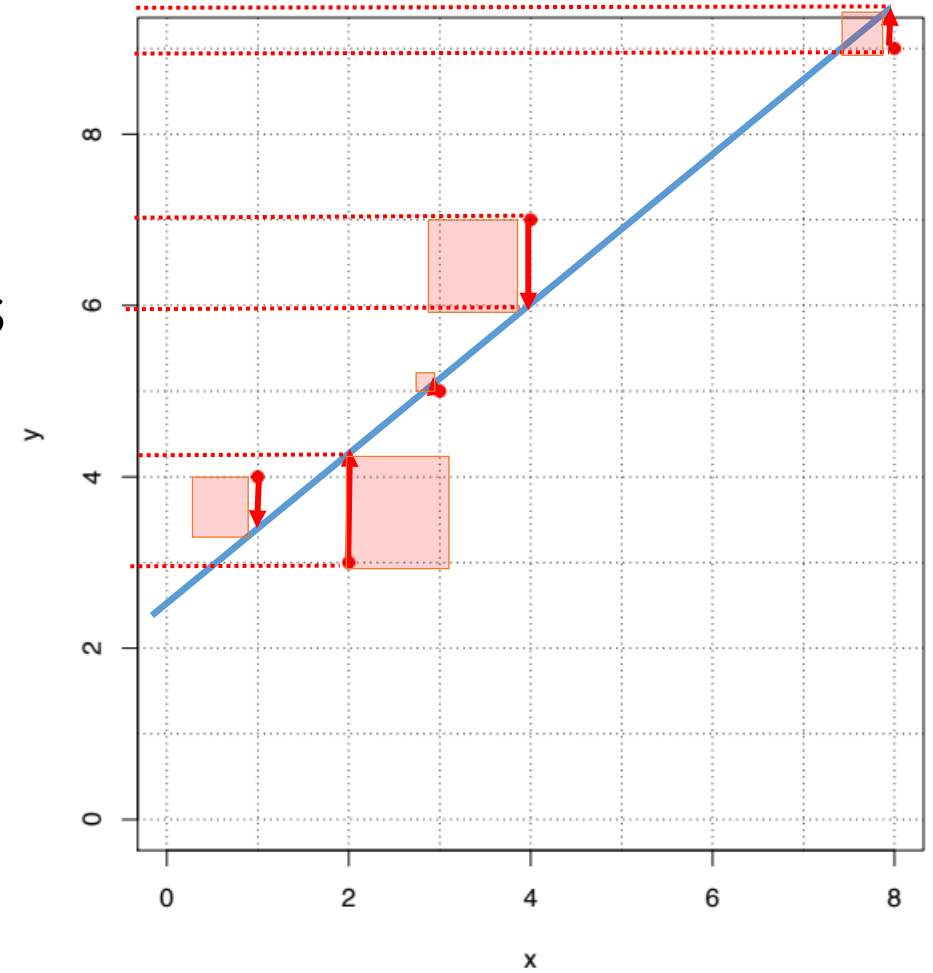
$$y_i = 2.2 + 0.83 x_i + \varepsilon_i$$

SS: 3.03

R^2

- Coefficient of determination
- Proportion of how much variance in y is explained by x

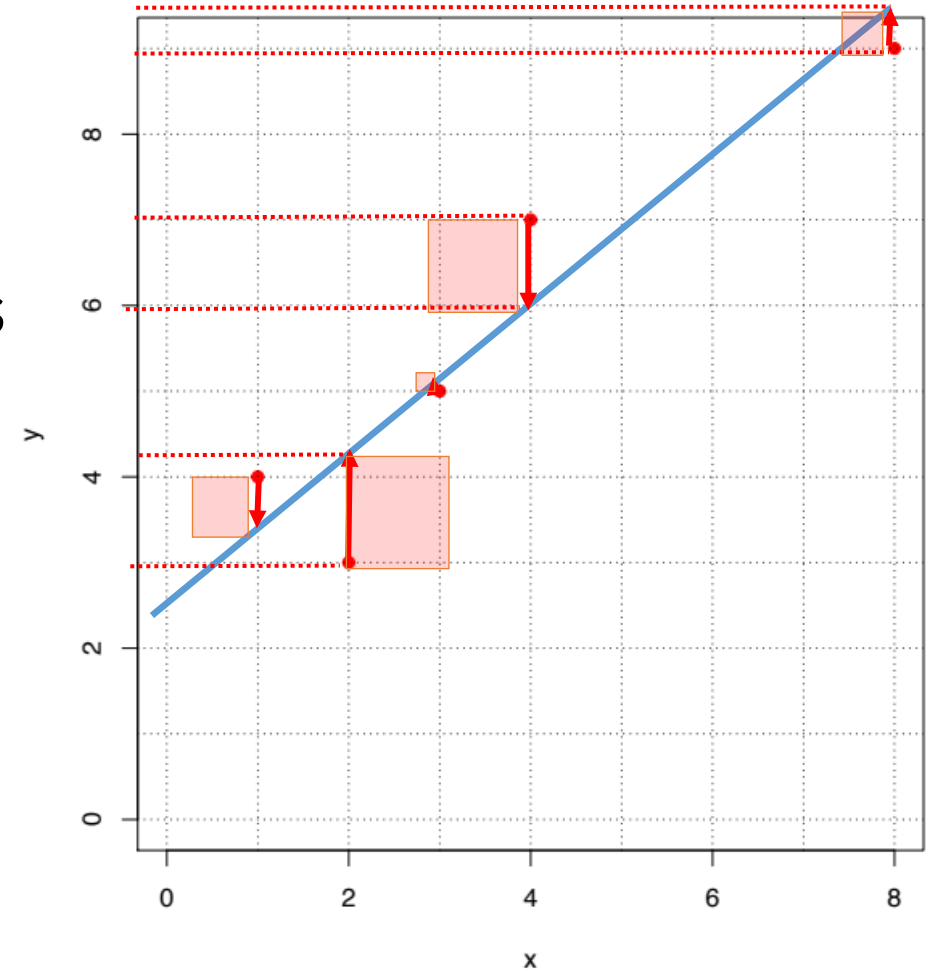
$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$



R^2

- Coefficient of determination
- Proportion of how much variance in y is explained by x

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

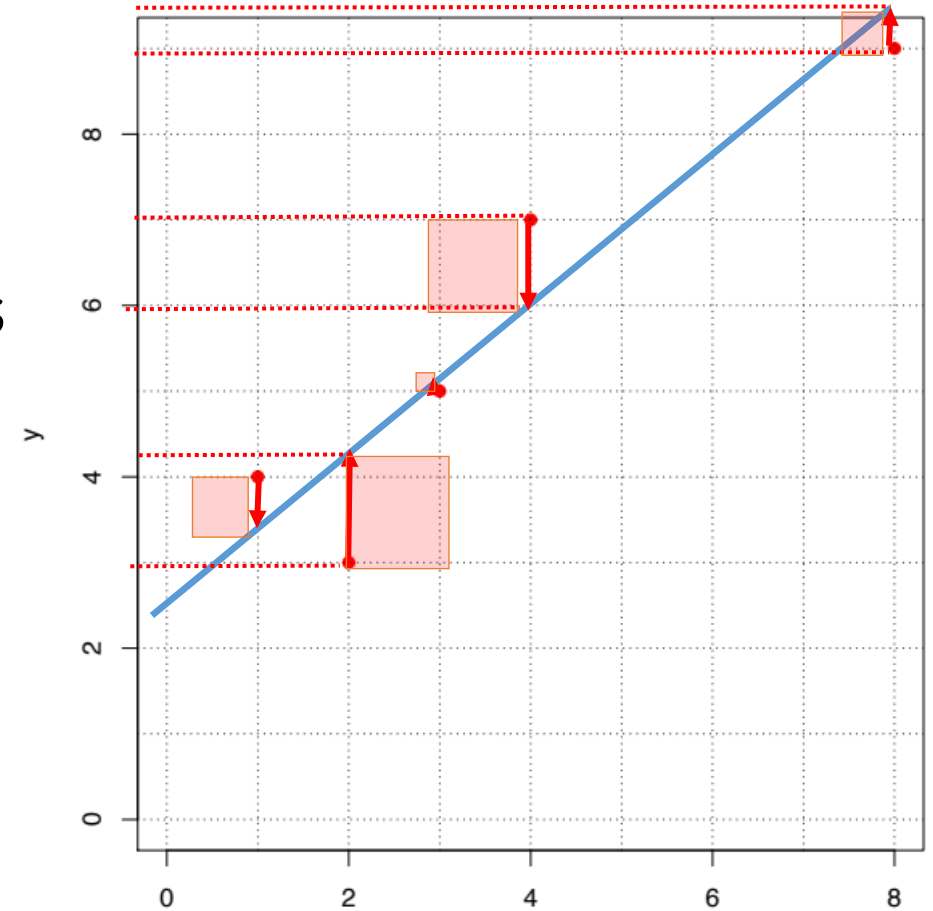


Wait, what?

R^2

- Coefficient of determination
- Proportion of how much variance in y is explained by x

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

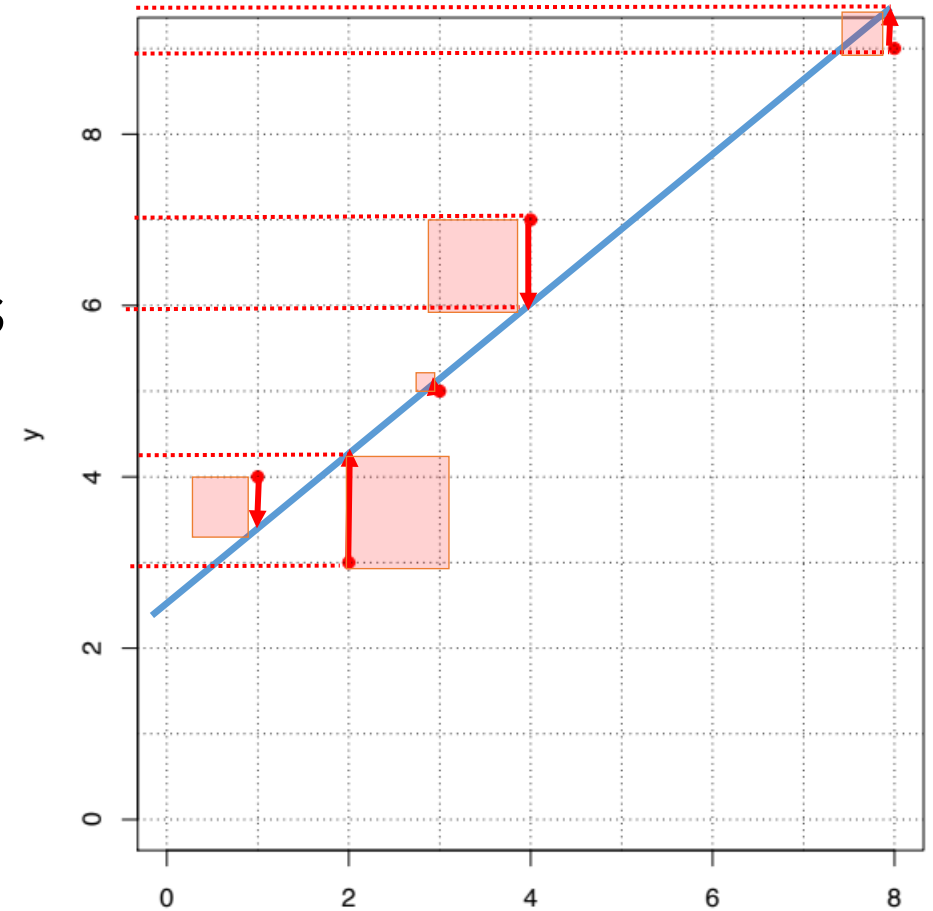


We know what SS_{res} is – the residual sum of squares. $\sum (y_i - \hat{y}_i)^2 = \sum (\epsilon_i)^2$

R^2

- Coefficient of determination
- Proportion of how much variance in y is explained by x

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$



We know what SS_{res} is – the residual sum of squares

$$\sum (y_i - \hat{y}_i)^2 = \sum (\epsilon_i)^2$$

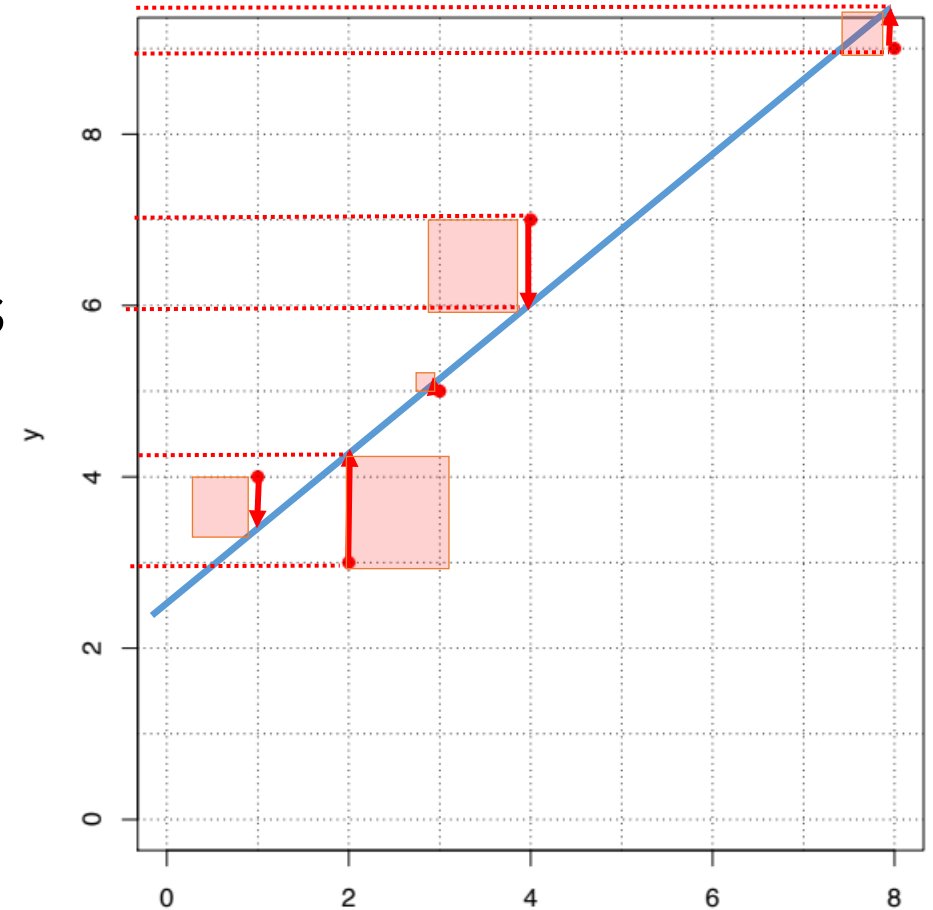
The total sum of squares is this:

$$\sum (y_i - \bar{y})^2$$

R^2

- Coefficient of determination
- Proportion of how much variance in y is explained by x

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$



We know what SS_{res} is – the residual sum of squares

$$\sum (y_i - \hat{y}_i)^2 = \sum (\epsilon_i)^2$$

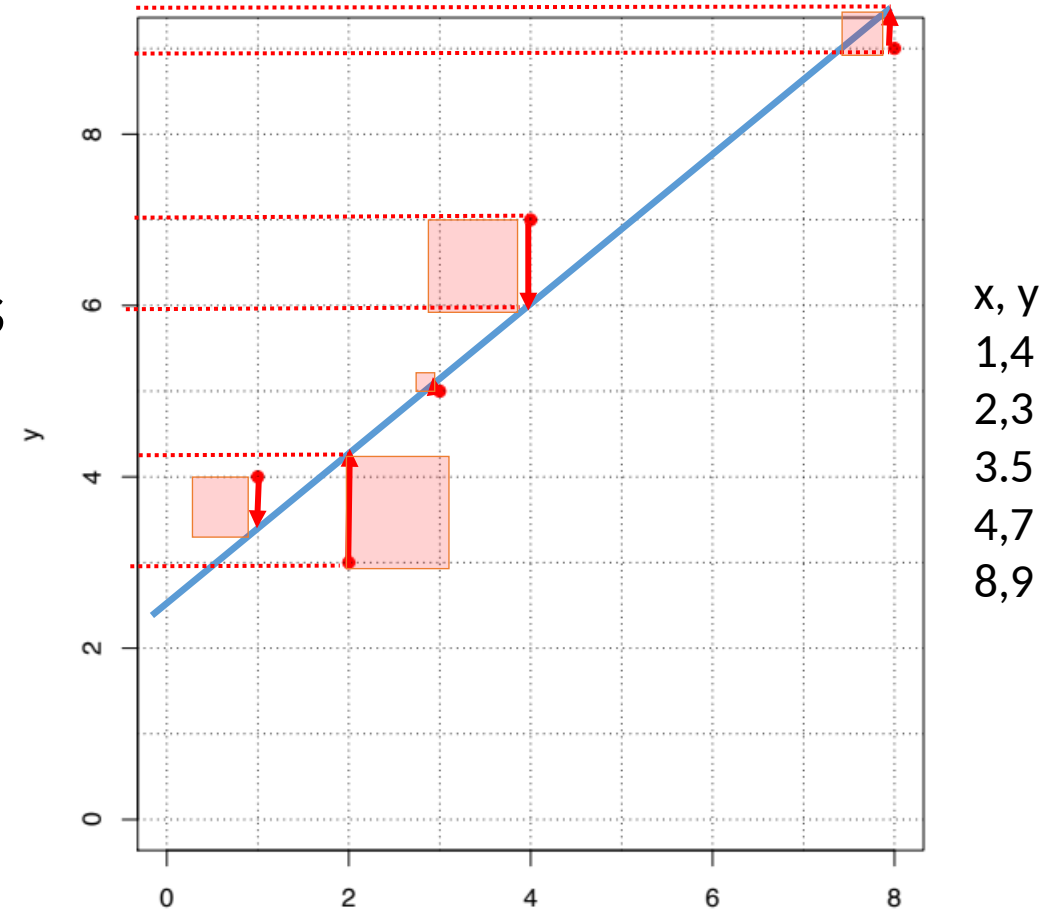
The total sum of squares is this:

$$\sum (y_i - \bar{y})^2 = \sigma^2 * (n - 1)$$

R²

- Coefficient of determination
- Proportion of how much variance in y is explained by x

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$



We know what SS_{res} is – the residual sum of squares

$$\sum (y_i - \hat{y}_i)^2 = \sum (\varepsilon_i)^2$$

3.03

The total sum of squares is this:

$$\sum (y_i - \bar{y})^2 = \sigma^2 * (n - 1)$$

5.8 * 4
23.2

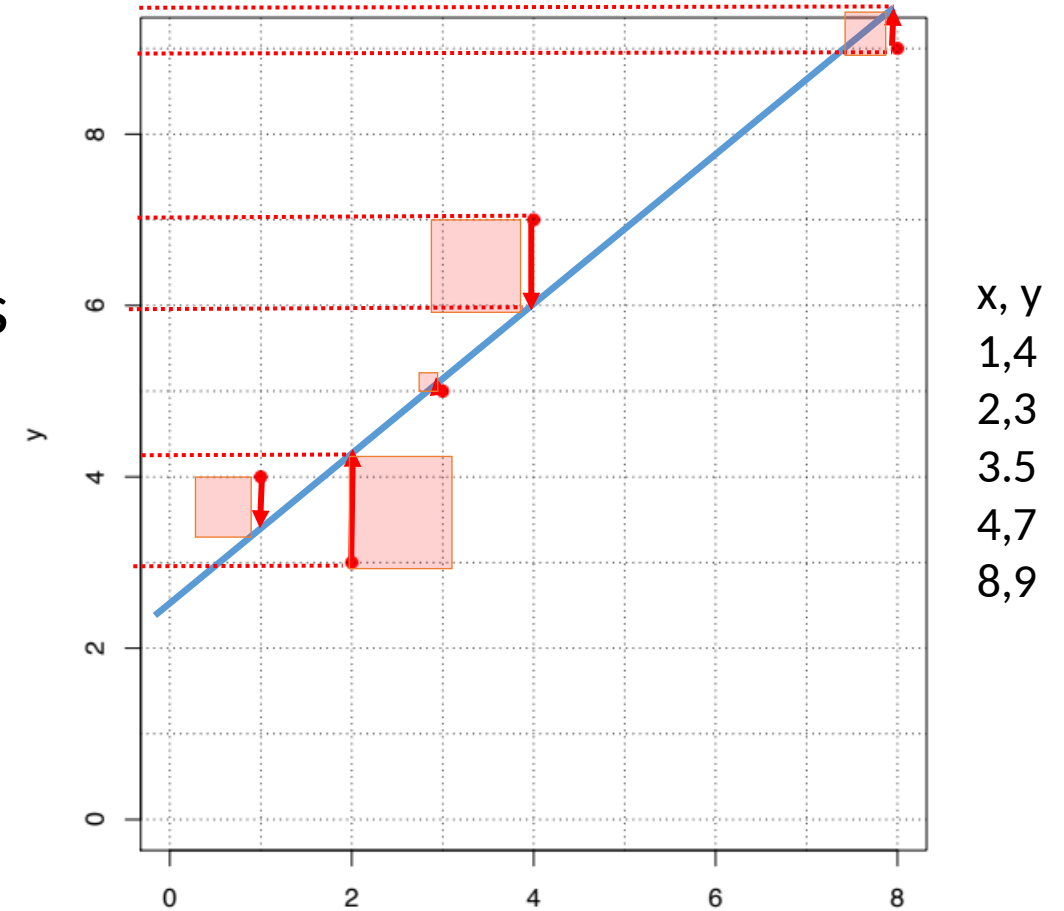
R²

- Coefficient of determination
- Proportion of how much variance in y is explained by x

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$= 1 - 3.03/23.2$$

$$= 0.87$$



We know what SS_{res} is – the residual sum of squares

$$\sum (y_i - \hat{y}_i)^2 = \sum (\varepsilon_i)^2$$

3.03

The total sum of squares is the nominator of :

$$\sum (y_i - \bar{y})^2 = \sigma^2 * (n - 1)$$

5.8 * 4
23.2

Linear regression:

- Minimizing sum of squared residuals of line
- Then get b_1 and b_0
- Calculate R^2 to assess how much variance in the response variable is explained by the explanatory variable

Exercise – no hand-out

$R^2 = 0.87$

-0.7
1.2
0.1
1
0.
SS: 3.03

- Run a linear regression in R with x and y as we've used them here.

```
x<-c(1, 2, 3, 4, 8)
```

```
y<-c(4, 3, 5, 7, 9)
```

```
model1 <- (lm(y~x))
```

```
model1
```

```
summary(model1)
```

```
anova(model1)
```

```
resid(model1)
```

```
cov(x, y)
```

```
var(x)
```

```
plot(y~x)
```

x, y

1,4

2,3

3.5

4,7

8,9