Statistics with SpaRRows

Lecture 9

Julia Schroeder

Outline

- Linear models

Linear models

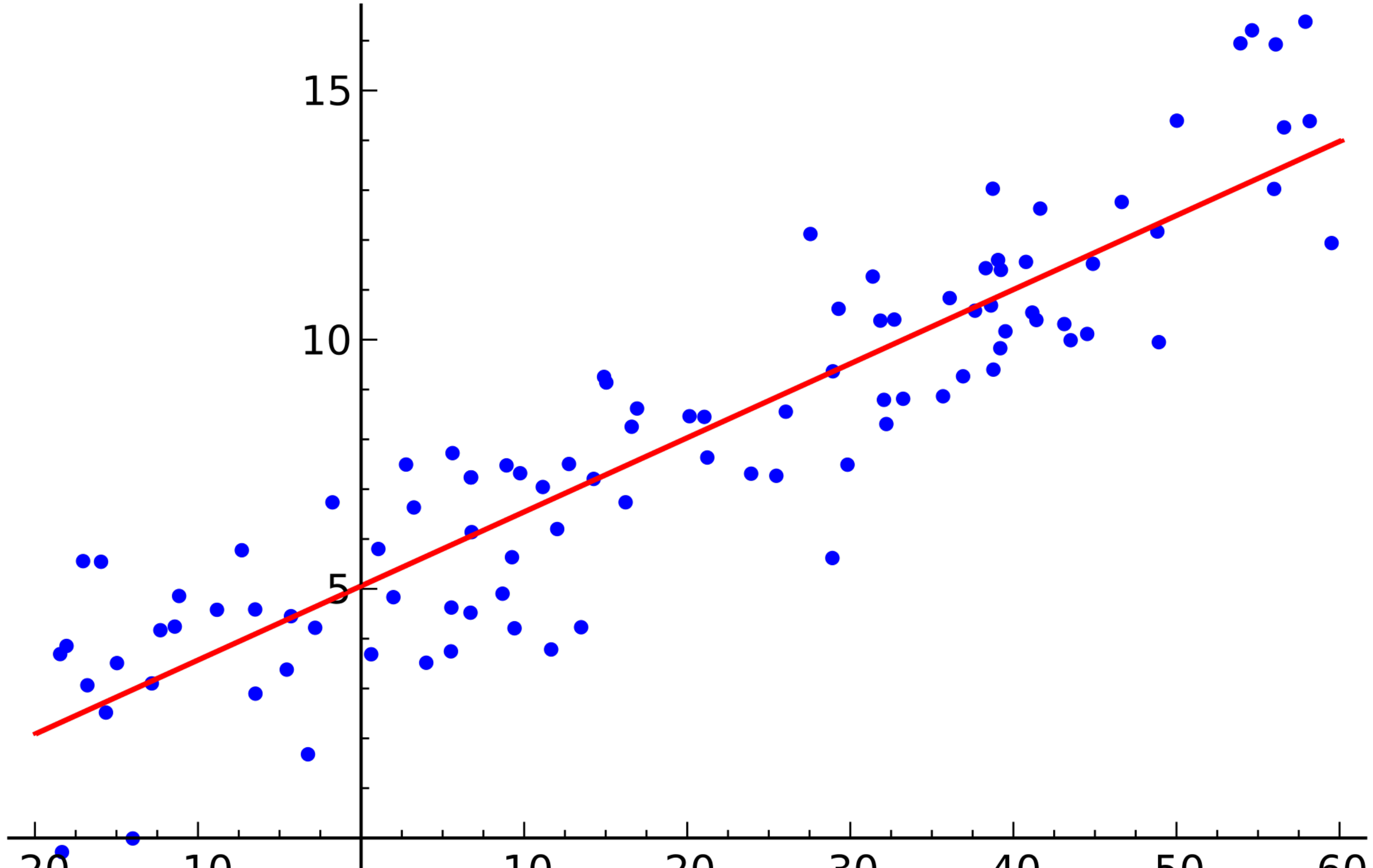
- Most important concepts
- Can model many questions (including previous t-test)
- If you understand linear models, everything else will be easy as a breeze!
- Aim to fit models to data

Fitting models to data

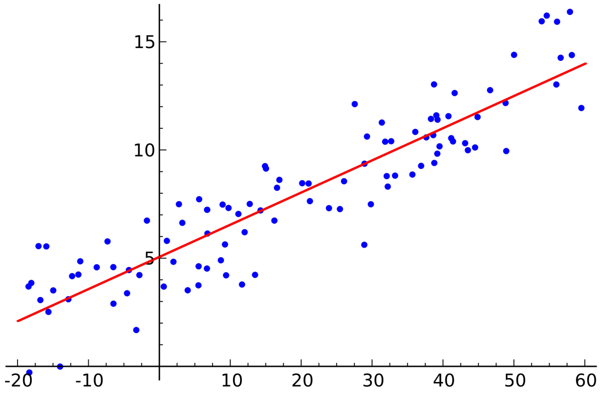
Fitting models to data

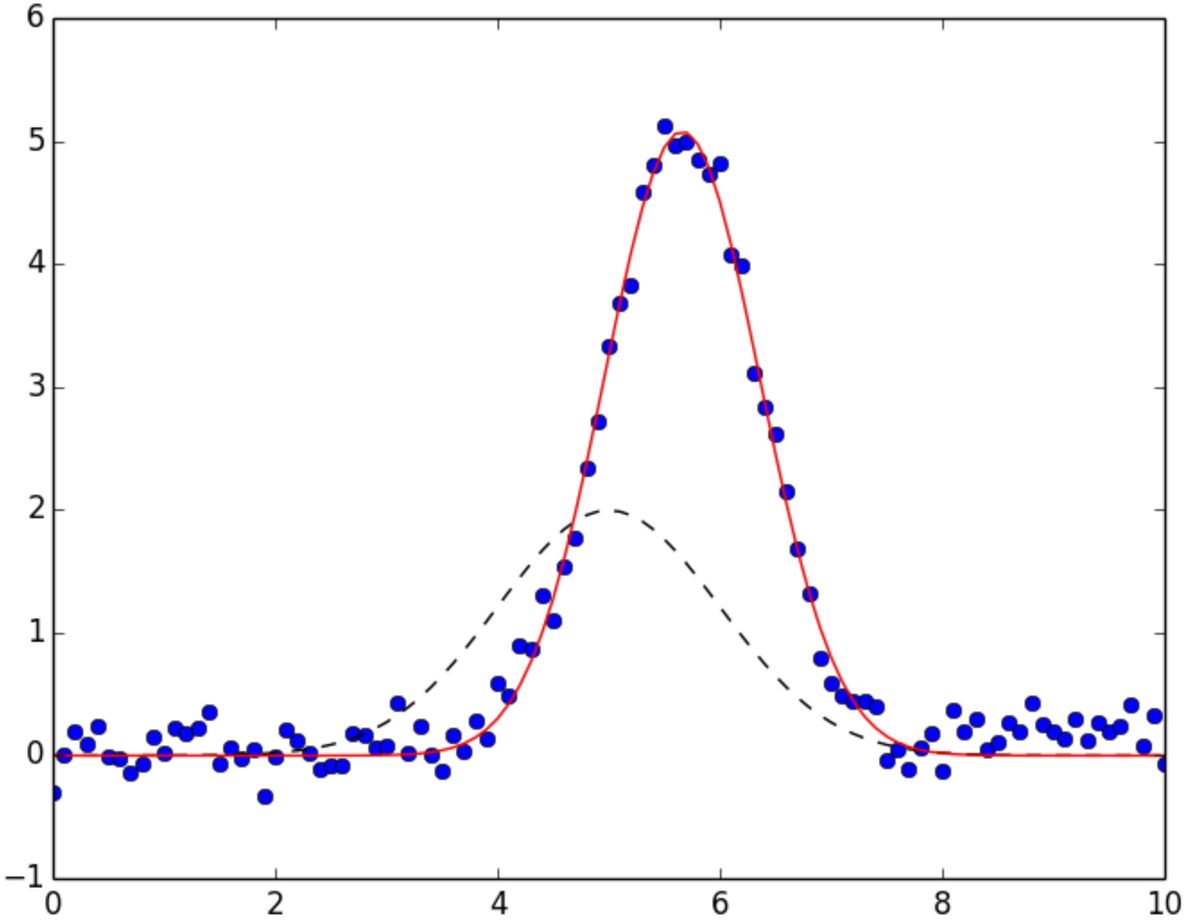
?

Fitting models to data

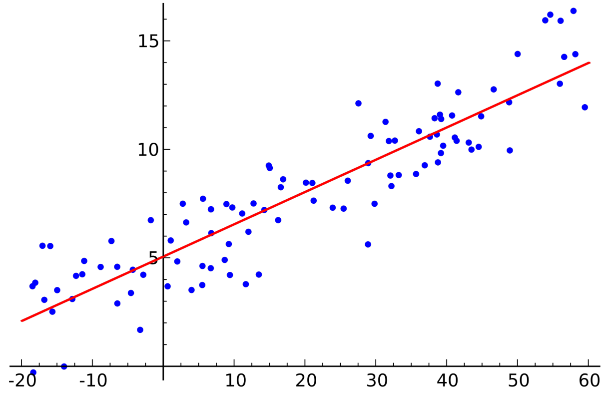


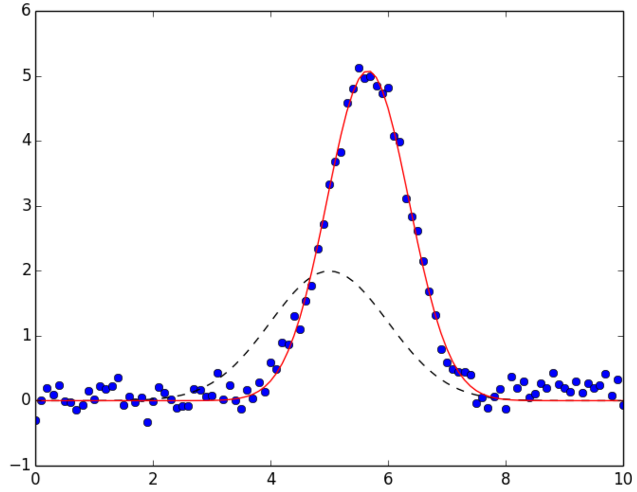
Fitting models to data

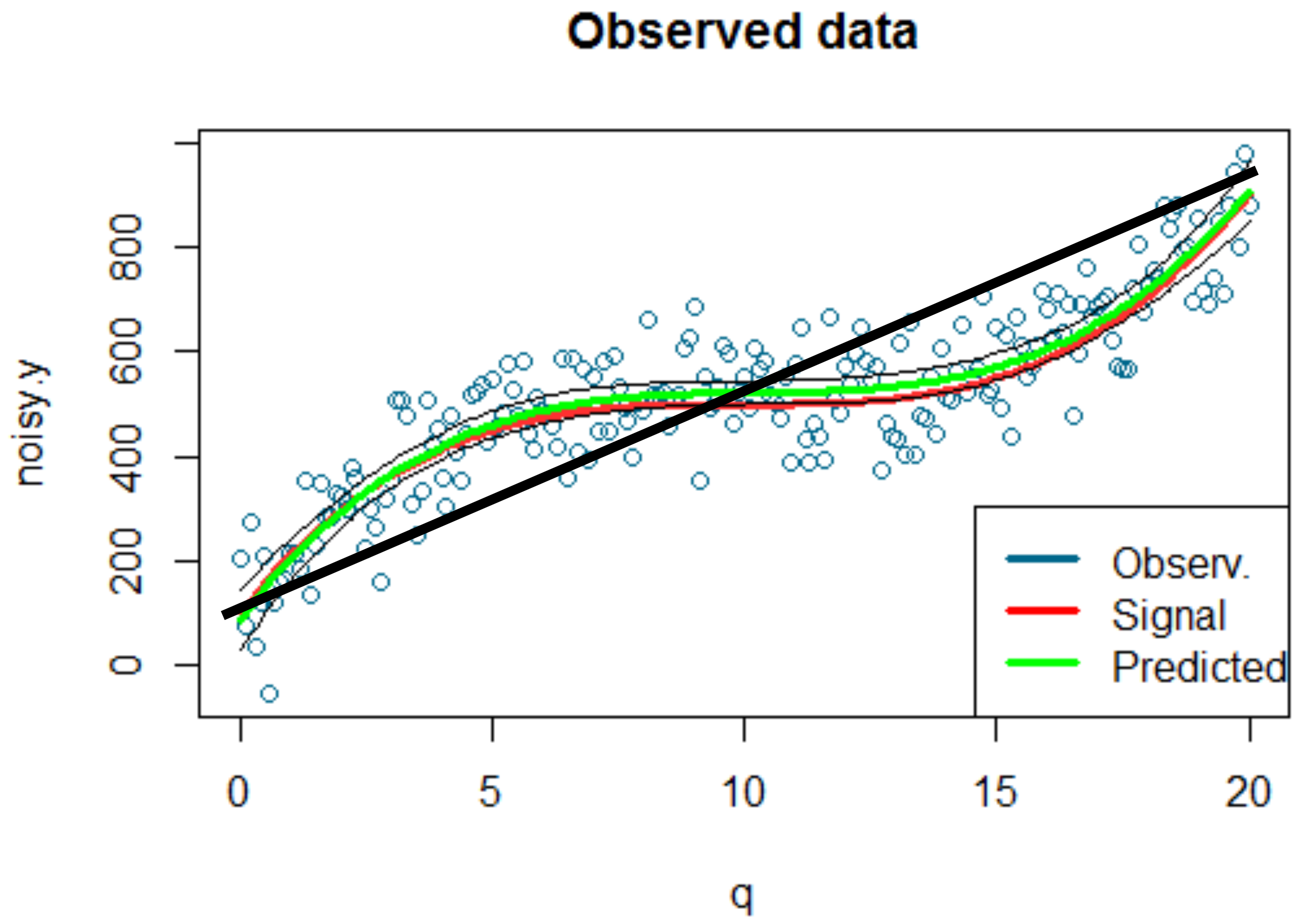




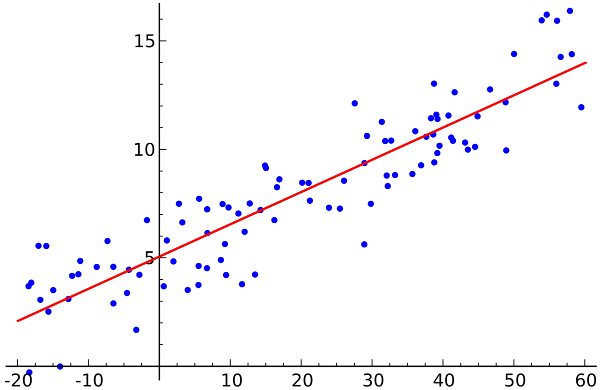
Fitting models to data

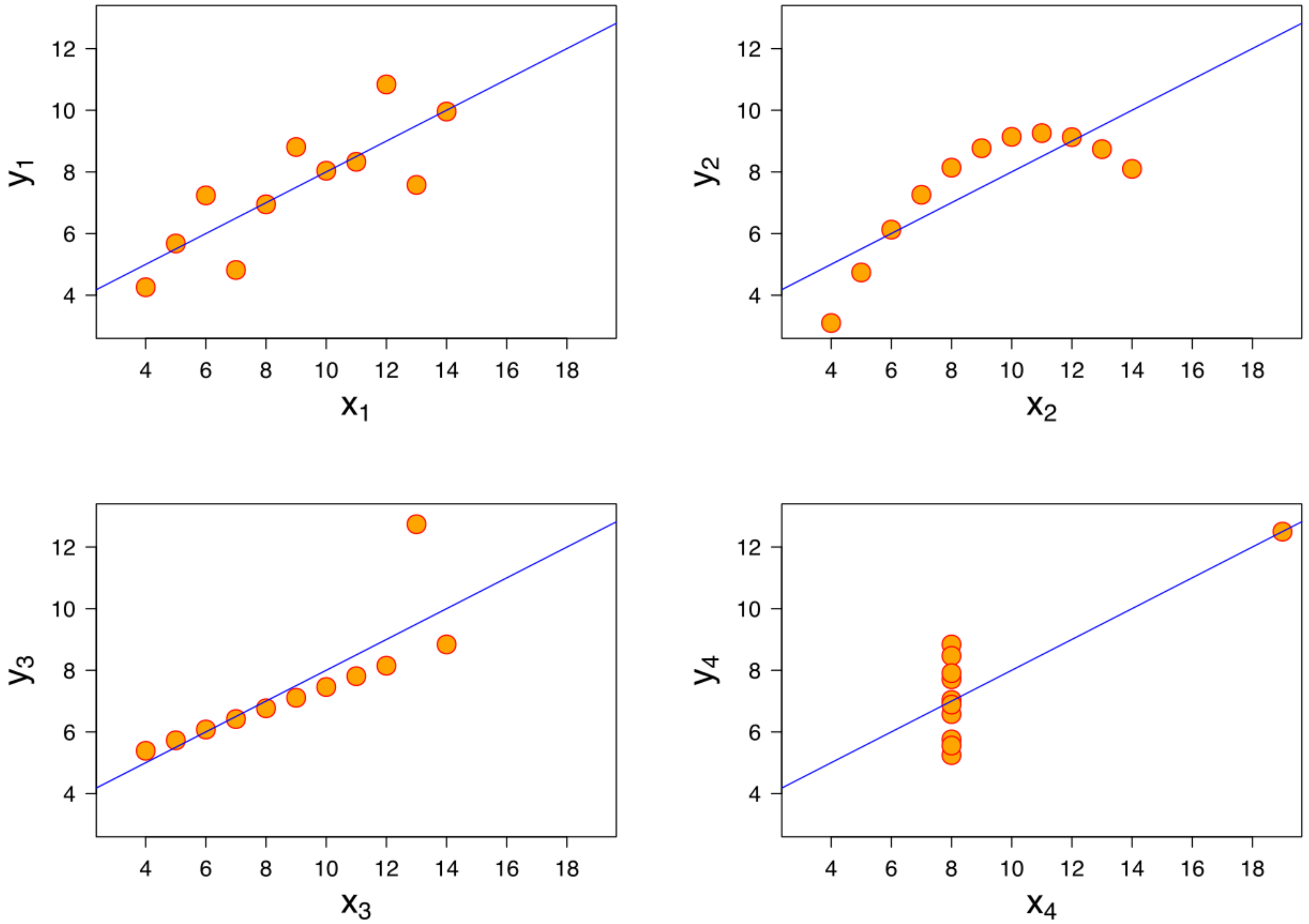






Fitting models to data





Linear models

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

Linear models

$$\color{red}{y_i} = b_0 + b_1 x_i + \varepsilon_i$$

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

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

ε	i
?	1
?	2
?	3
?	4
?	5

Linear models

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

$b_1 = ?$

ε	i
?	1
?	2
?	3
?	4
?	5

Linear models

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

$b_0 = ?$

$b_1 = ?$

ε	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
3	1
1	2
4	3
8	4
2	5

ε	i
?	1
?	2
?	3
?	4
?	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 ε_i 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

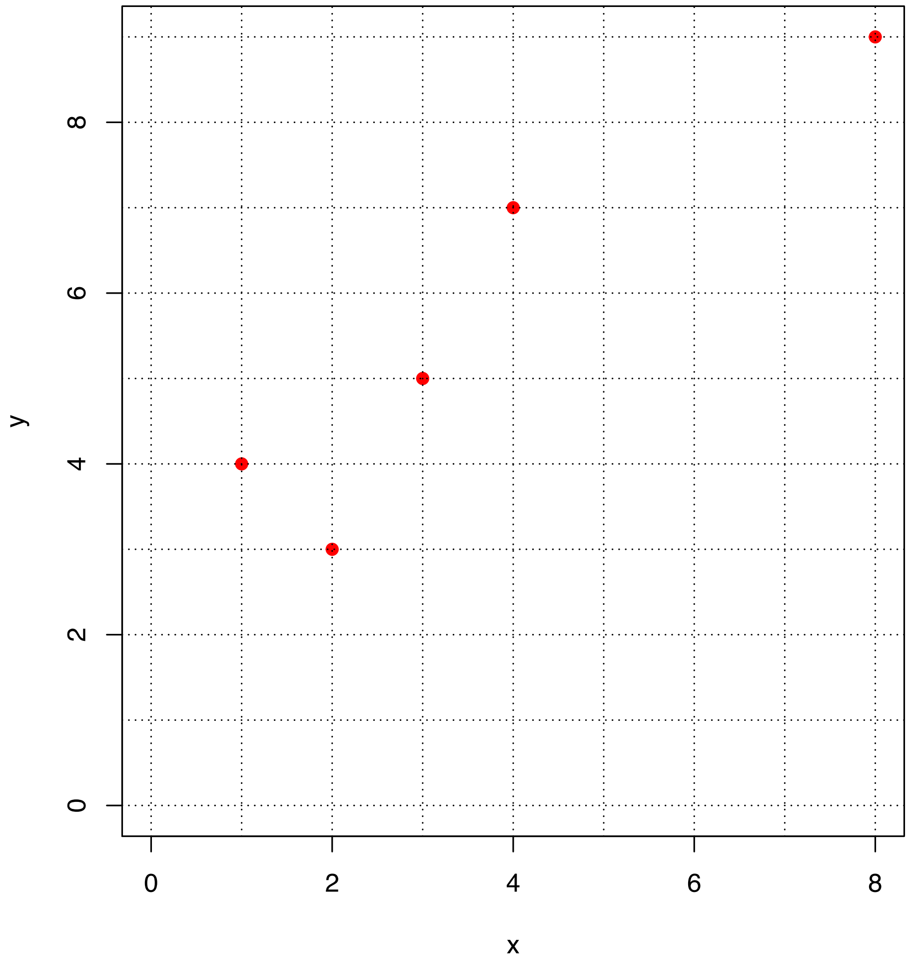
y	x
5	3
4	1
7	4
9	8
3	2

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

Let’s give this a try…

- Let’s plot this

y	x
5	3
4	1
7	4
9	8
3	2

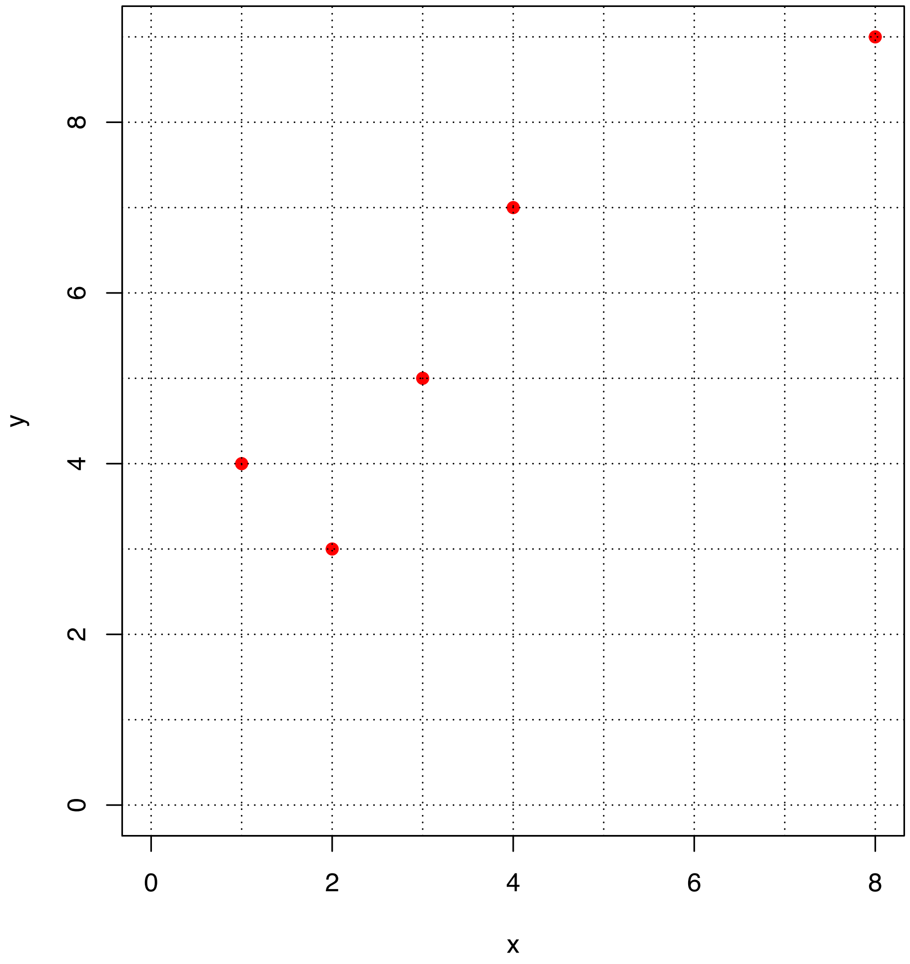


$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Let’s give this a try…

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

y	x
5	3
4	1
7	4
9	8
3	2

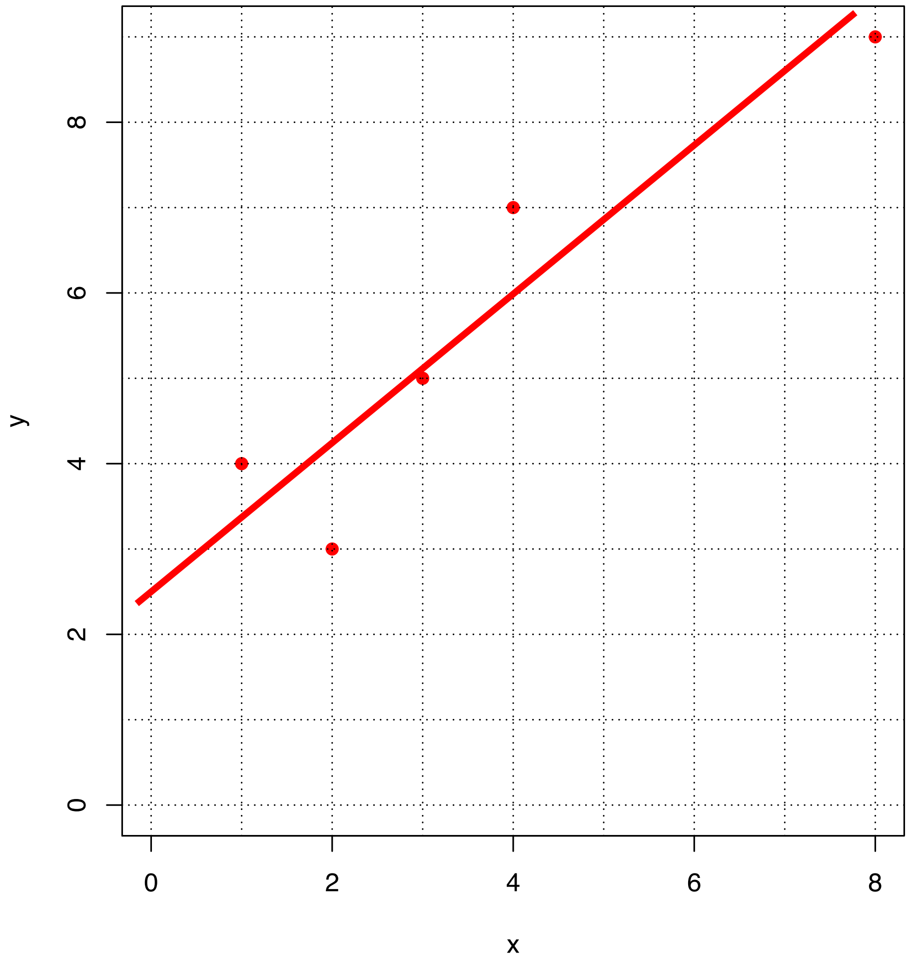


$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Let's give this a try…

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

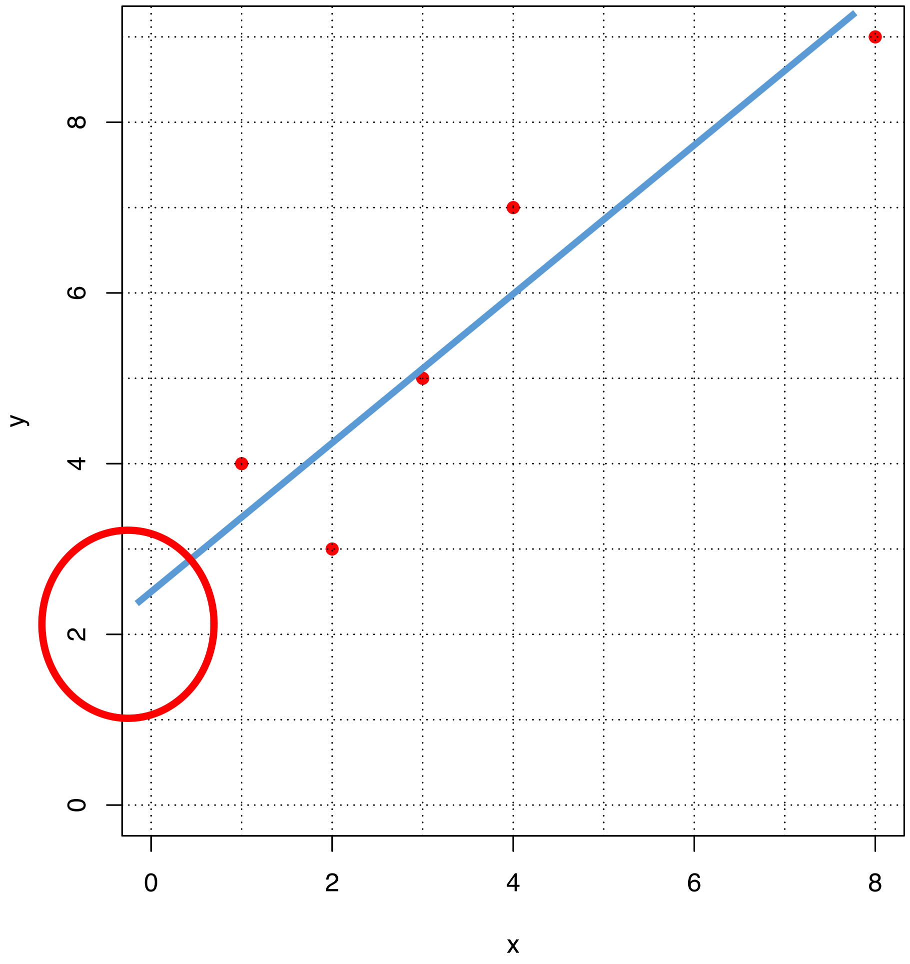
y	x
5	3
4	1
7	4
9	8
3	2



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

Let’s give this a try…

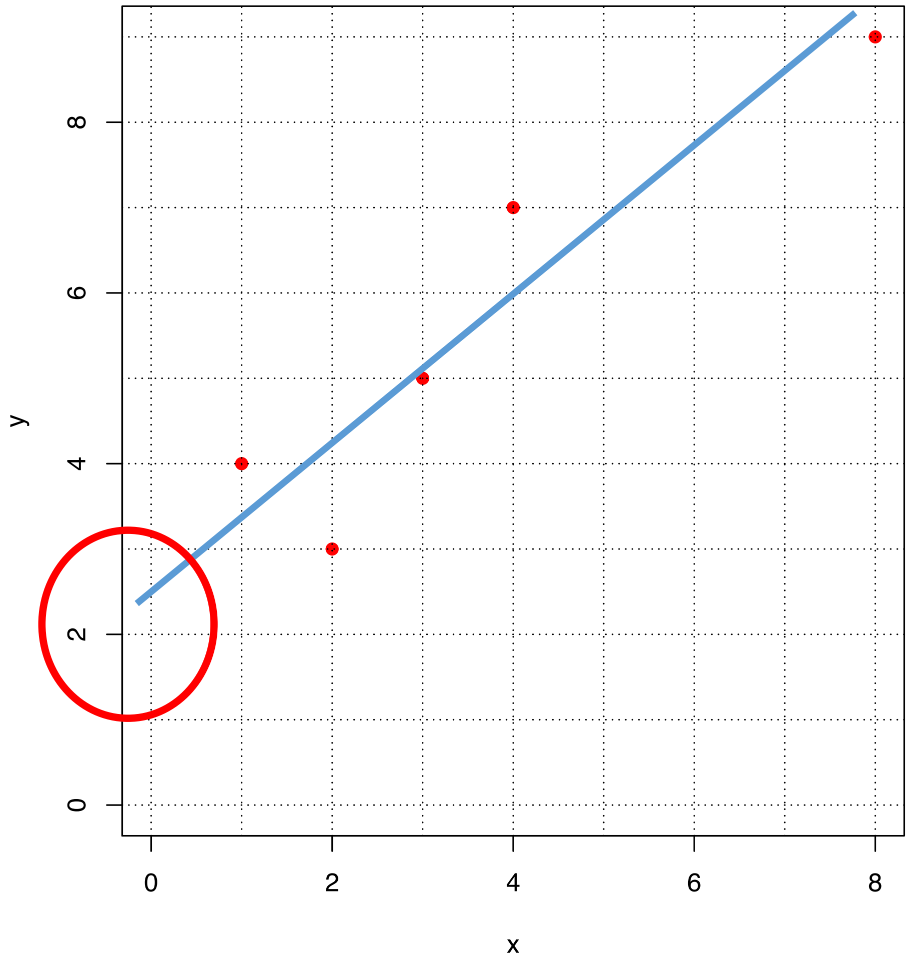
- Let’s plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1:
- Intercept:



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

Let's give this a try…

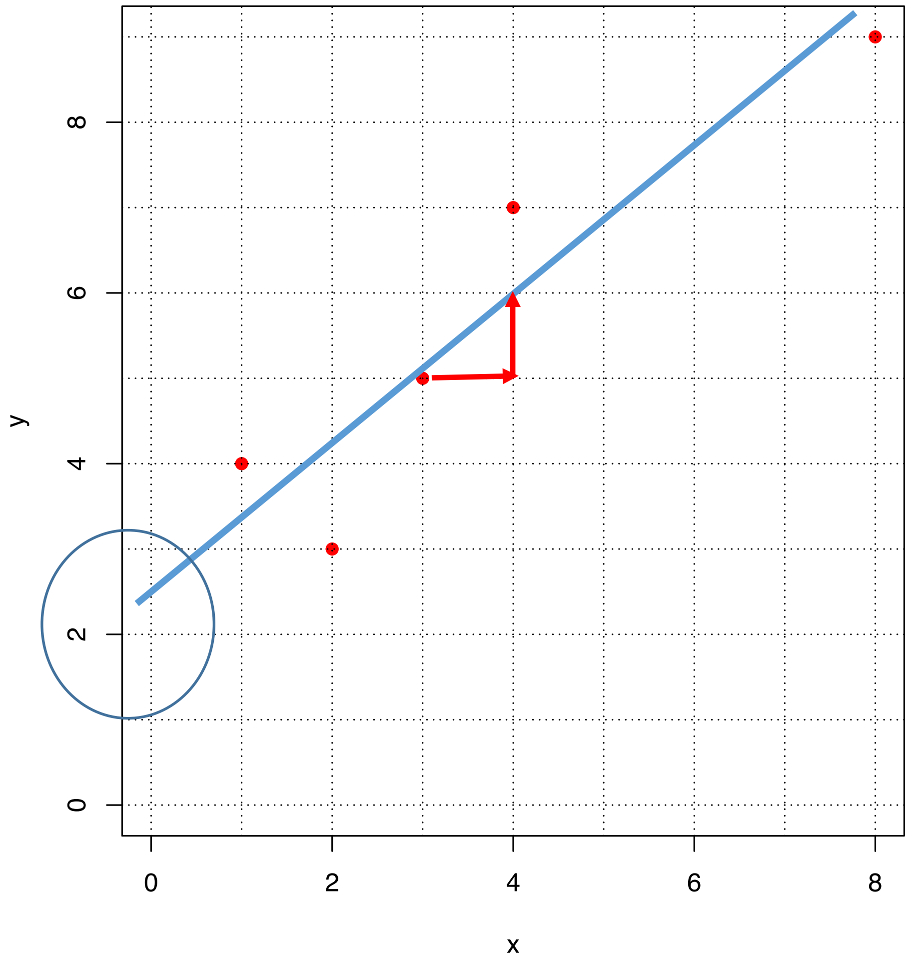
- Let's plot this
- Now we "guesstimate" the line
- Now we "guesstimate" b_0 and b_1:
- Intercept: something 2.2



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

Let’s give this a try…

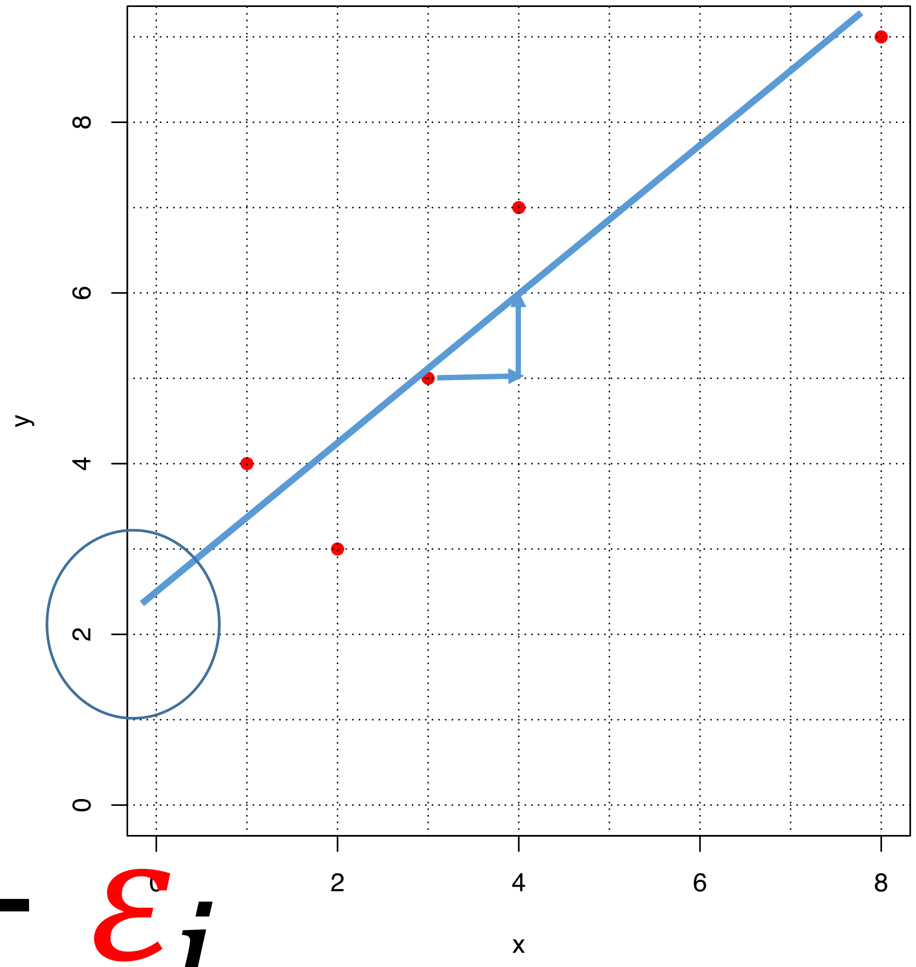
- Let’s plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1:
- Intercept: something 2.2
- Slope: close enough to 1



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

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

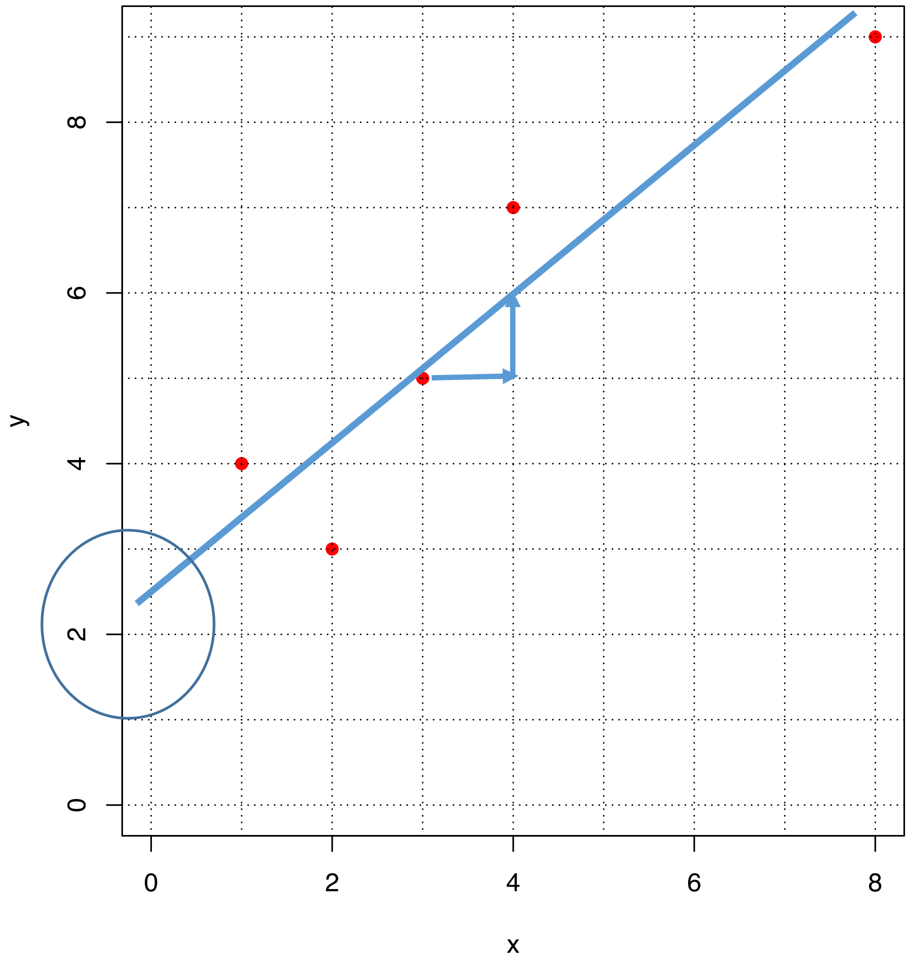


$$y_i = 2.2 + 1x_i + \varepsilon_i$$

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

Let’s give this a try…

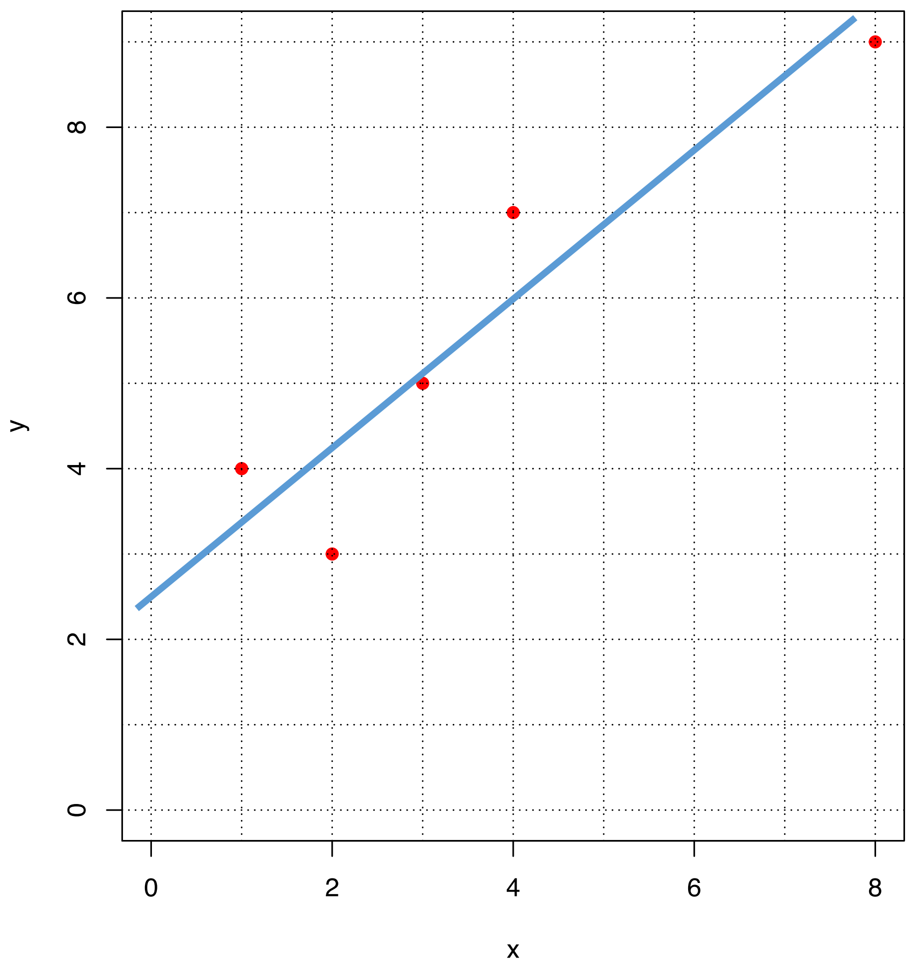
- Let’s plot this
- Now we “guesstimate” the line
- Now we “guesstimate” b_0 and b_1:
- Intercept: something 2.2
- Slope: close enough to 1
- But what’s with ε_i?



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

Let’s give this a try…

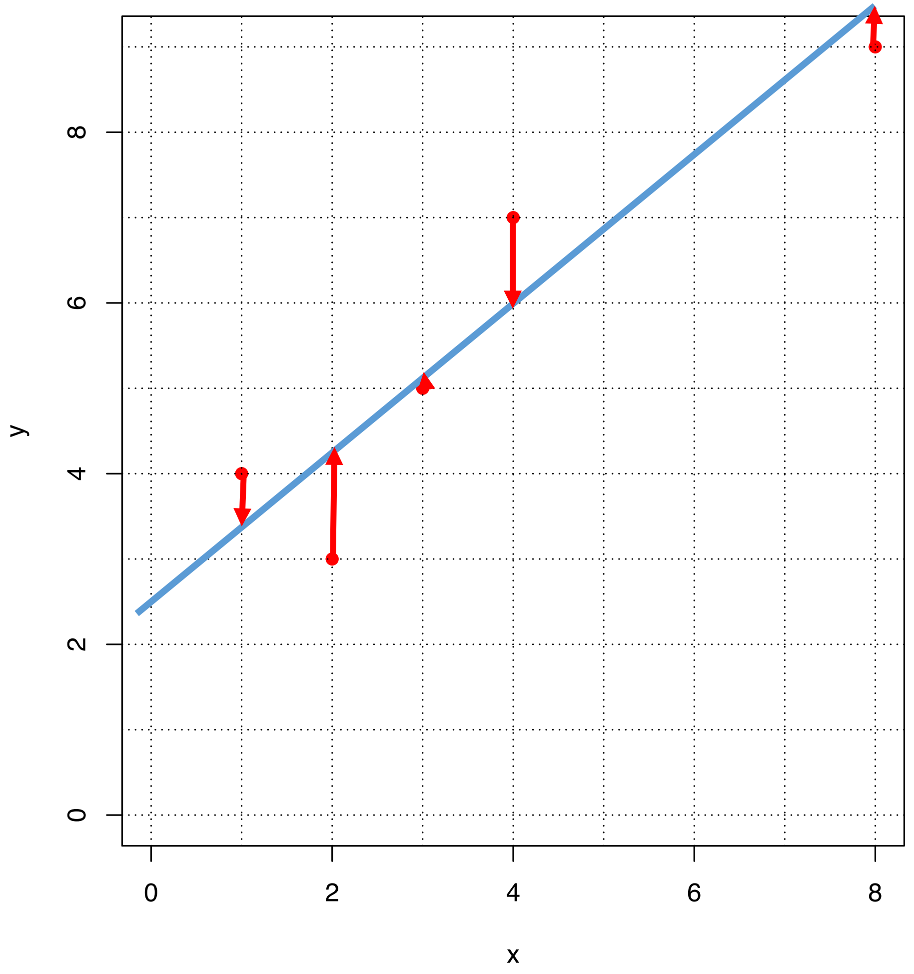
- But what’s with ε_i?
- The residuals are the “error” of the model
- We get them by plotting the vertical (y) distance:



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

Let's give this a try…

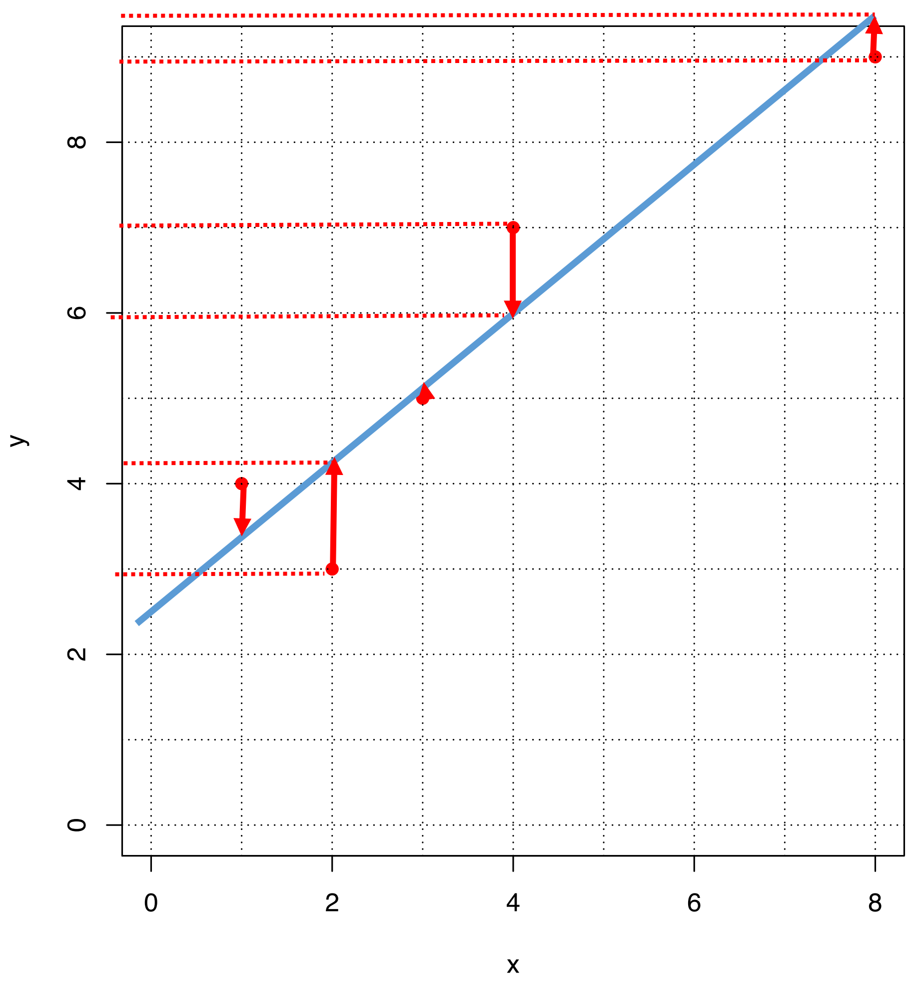
- But what's with ε_i?
- The residuals are the "error" of the model
- We get them by plotting the vertical (y) distance:



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

Let's give this a try…

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

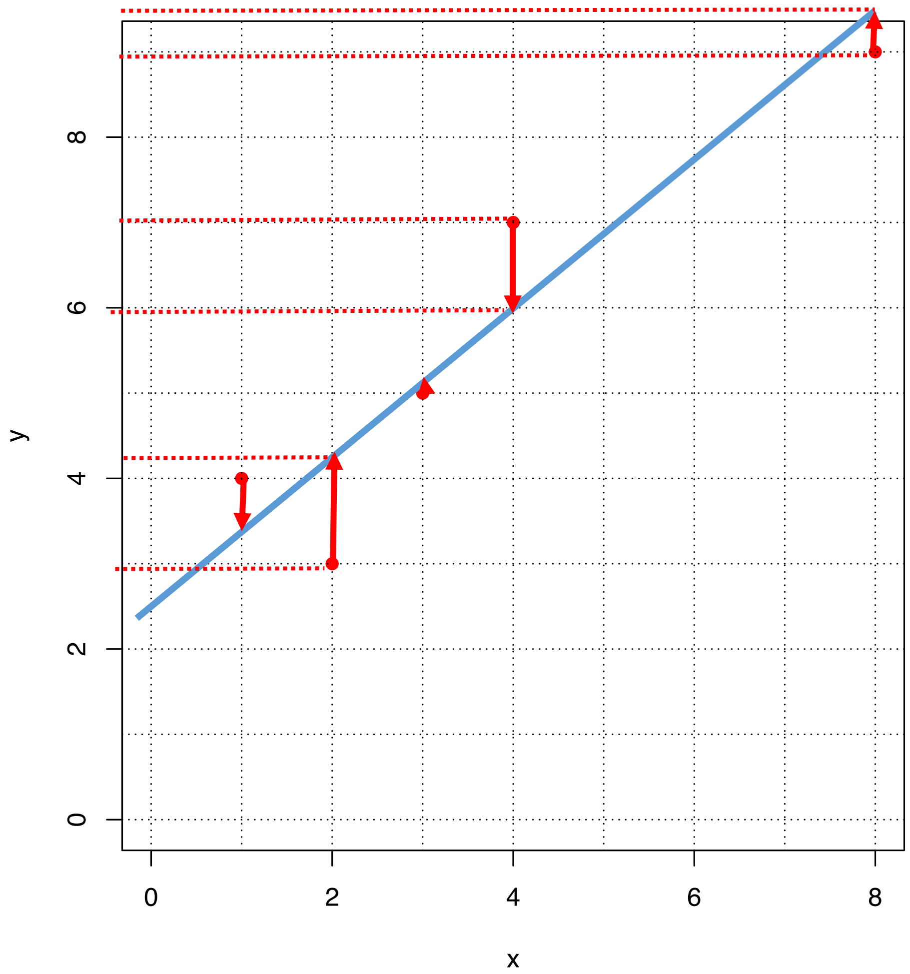


ε
-0.7
1.2
0.1
1
0.3

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

Let’s give this a try…

- But what’s with ε_i?
- 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?

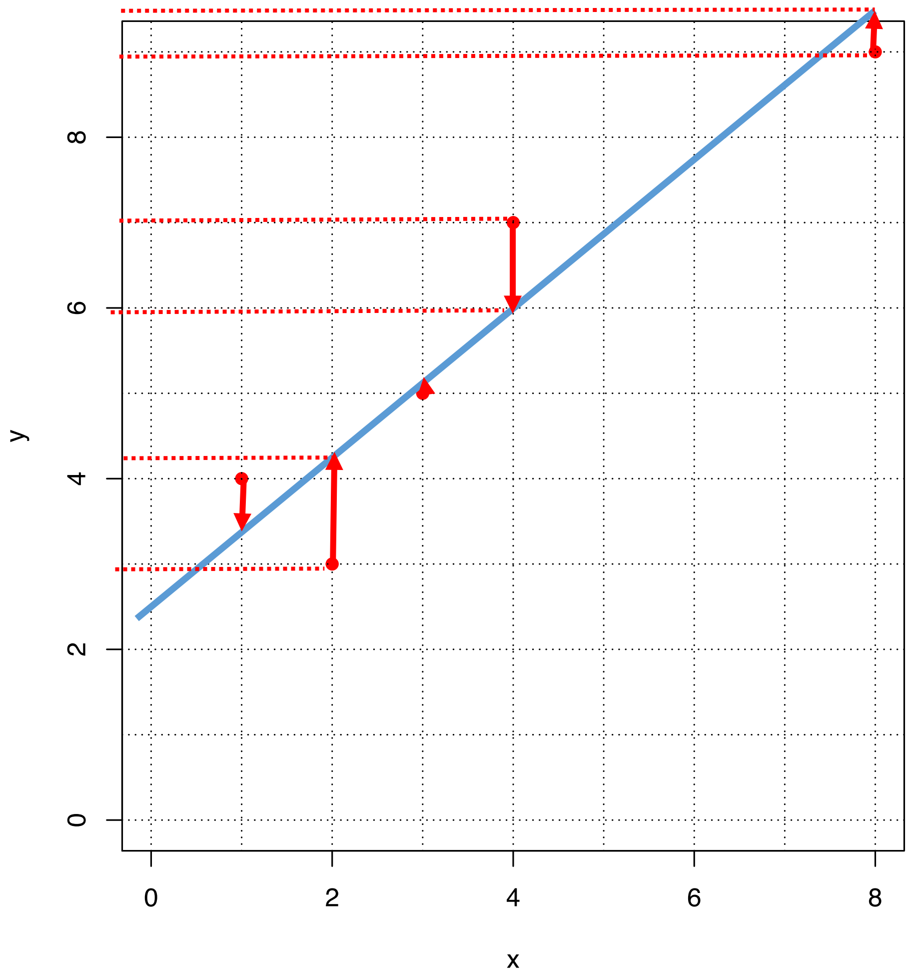


$$y_i = 2.2 + 1x_i + \varepsilon_i$$

ε
-0.7
1.2
0.1
1
0.3

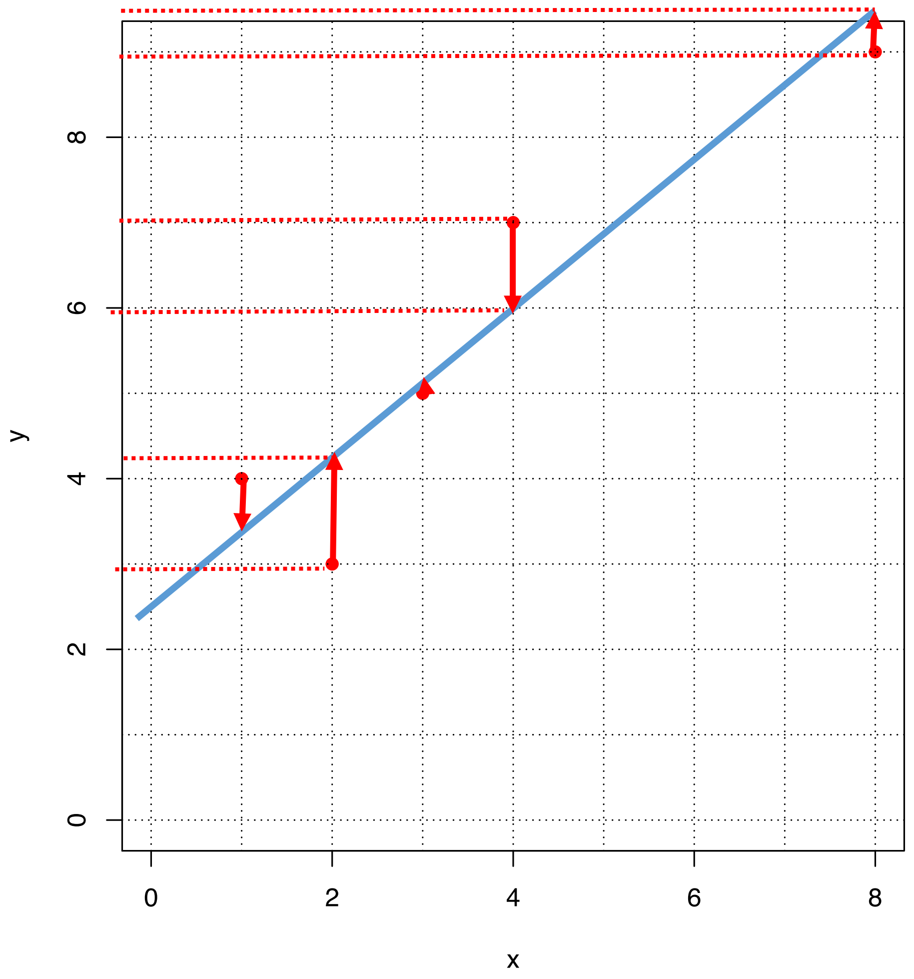
No more guesstimates…

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

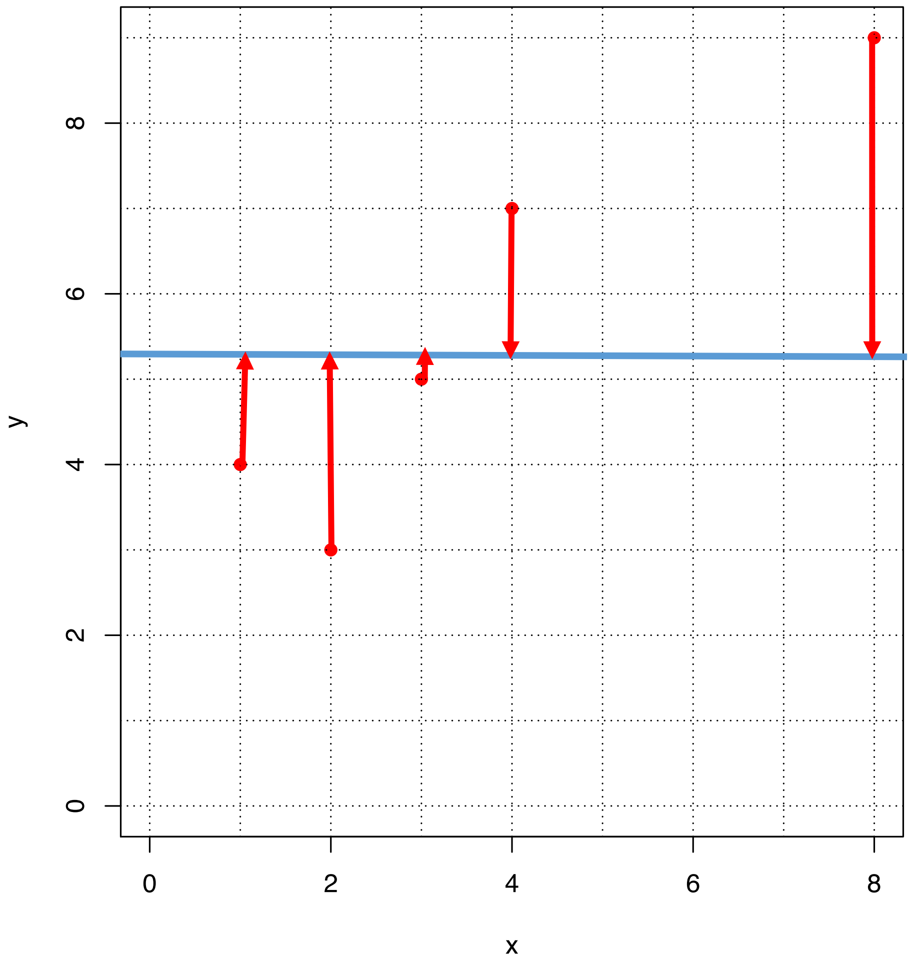


No more guesstimates…

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



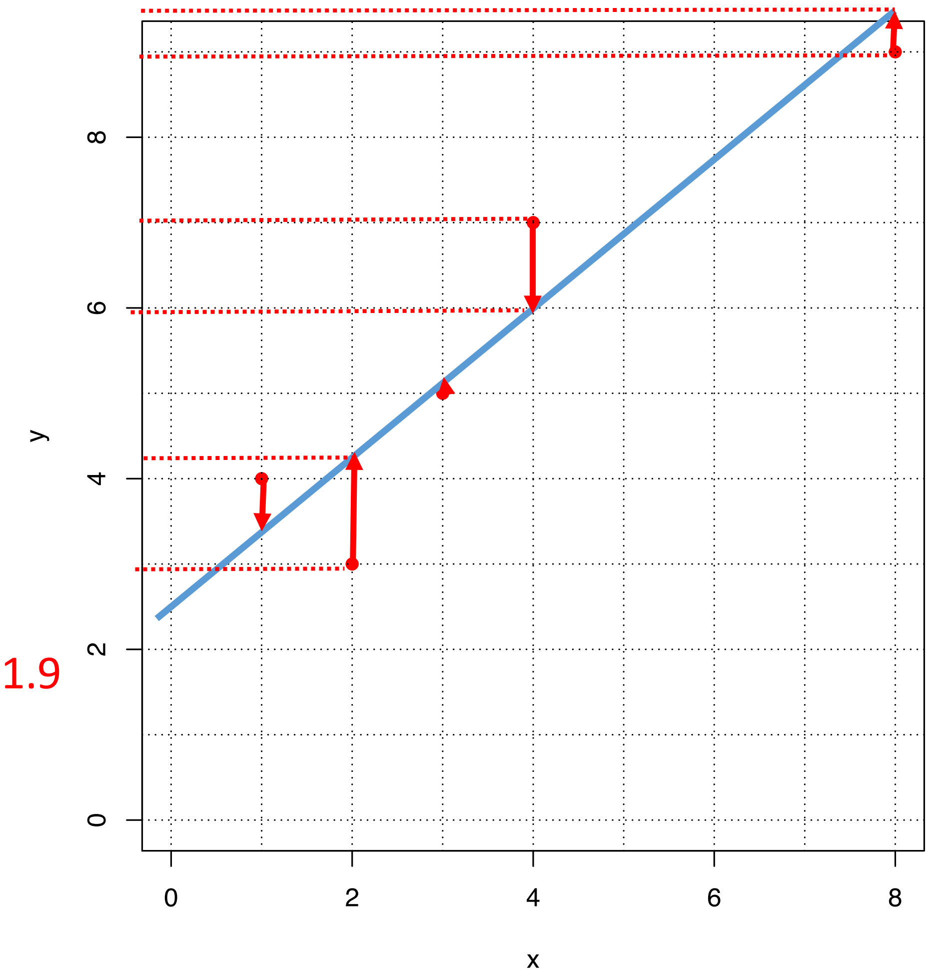
No more guesstimates…



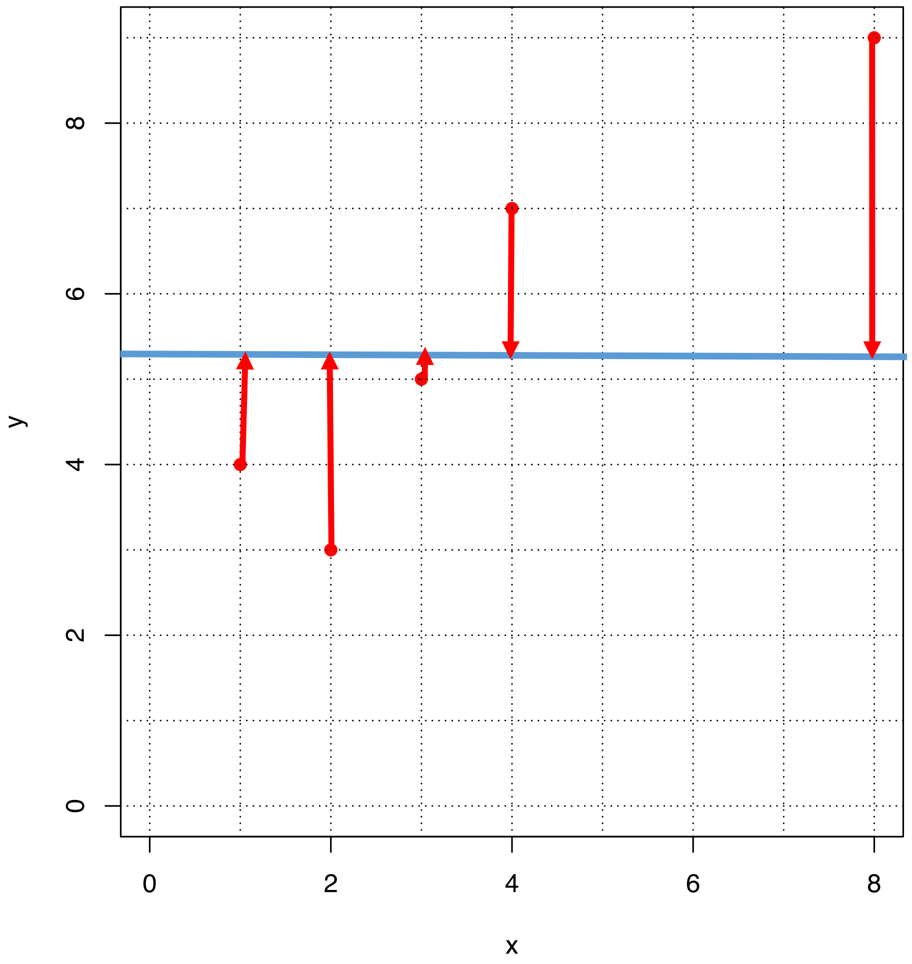
ε

-0.7
1.2
0.1
1
0.3

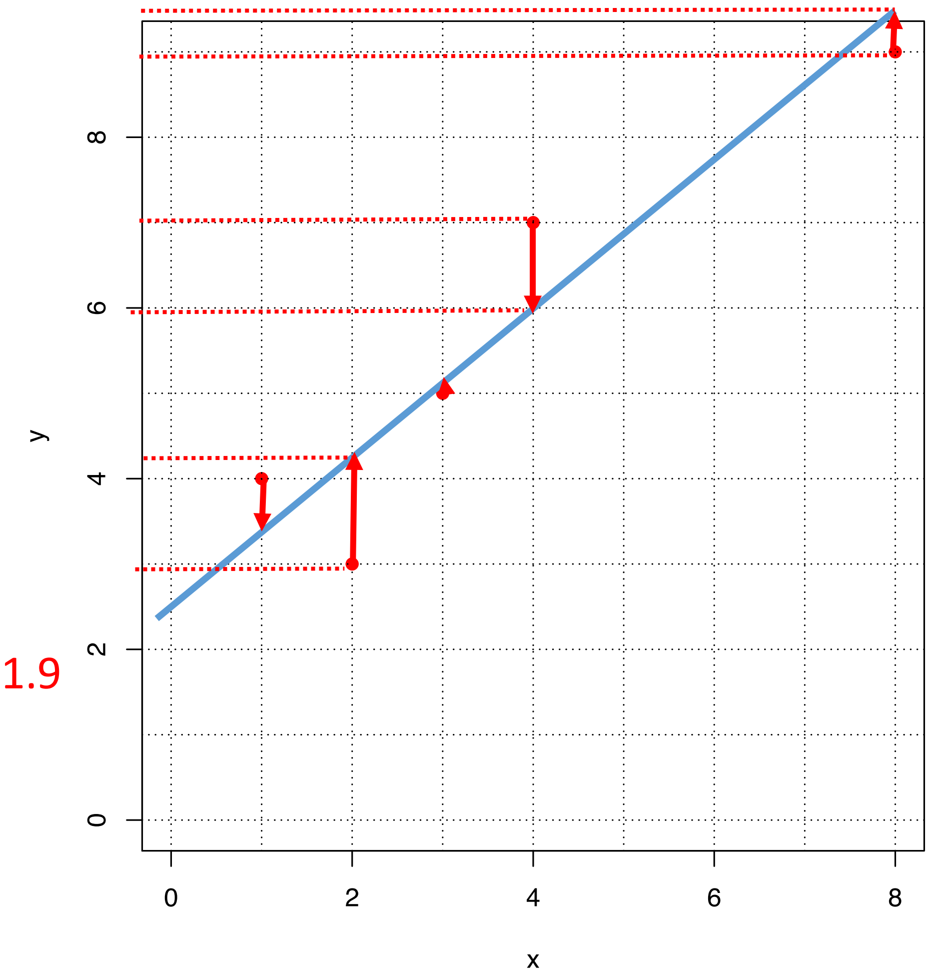
Sum: 1.9



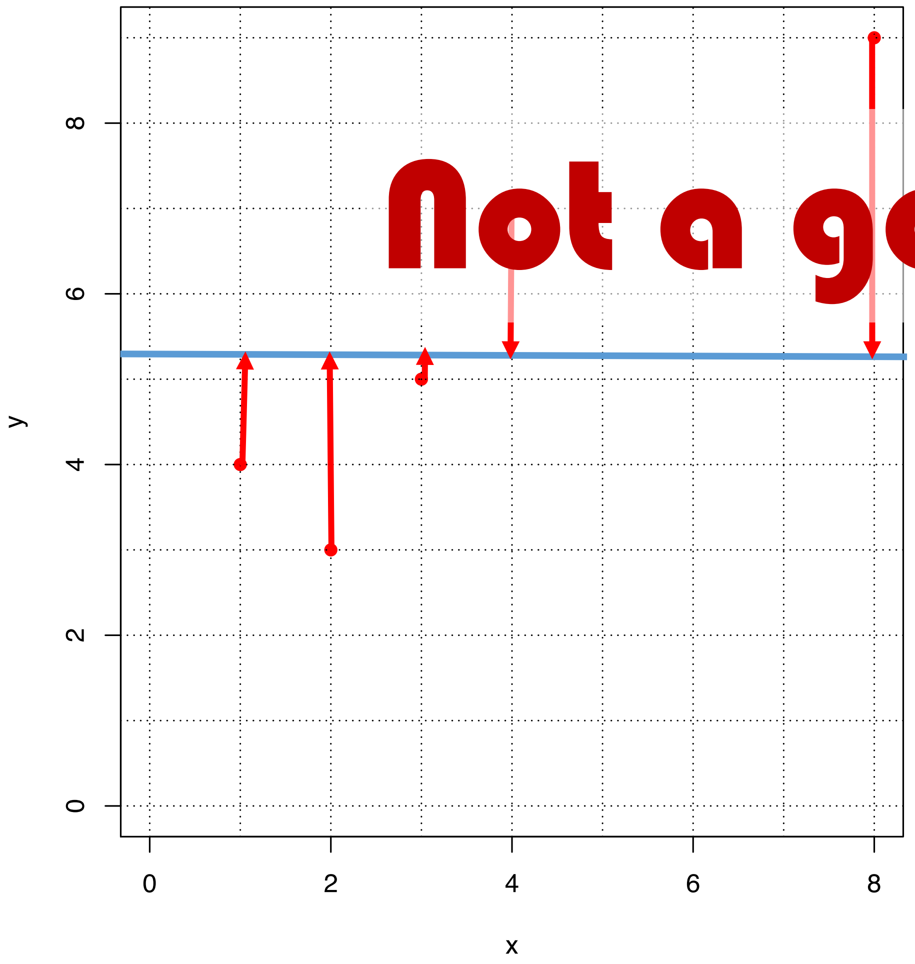
No more guesstimates...



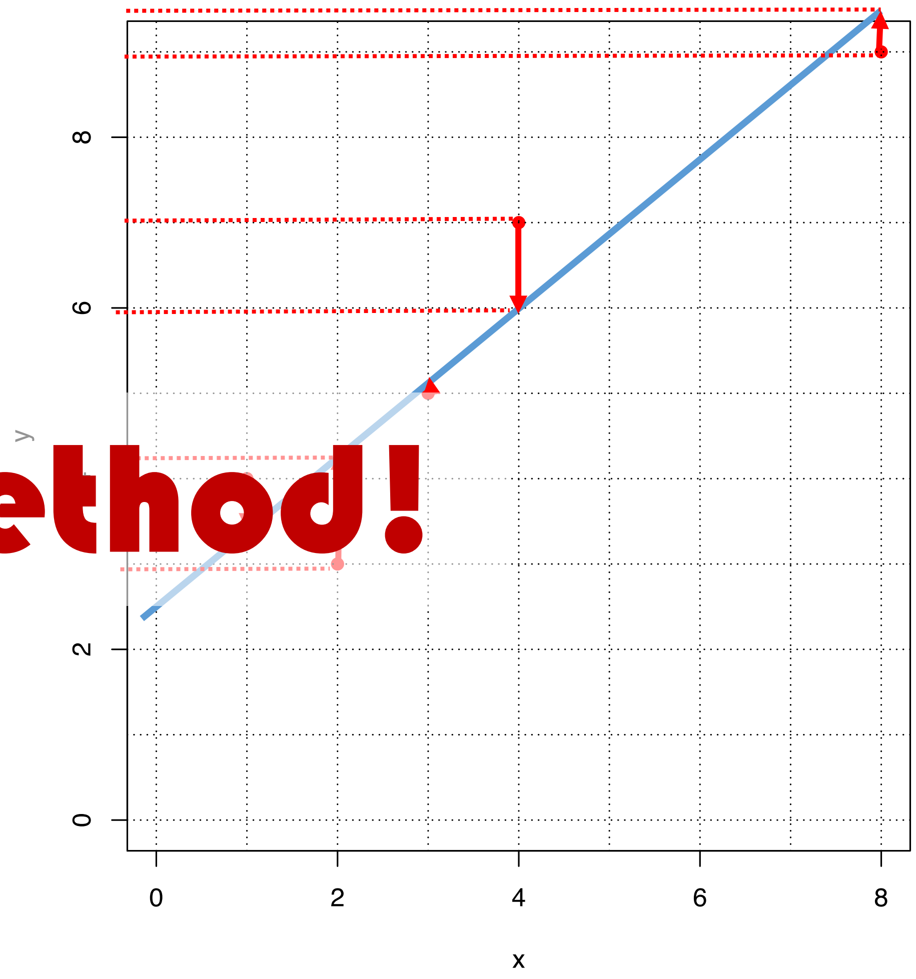
ε	ε
1.2	-0.7
2.2	1.2
0.2	0.1
-1.8	1
-3.4	0.3
Sum: -1.6	Sum: 1.9



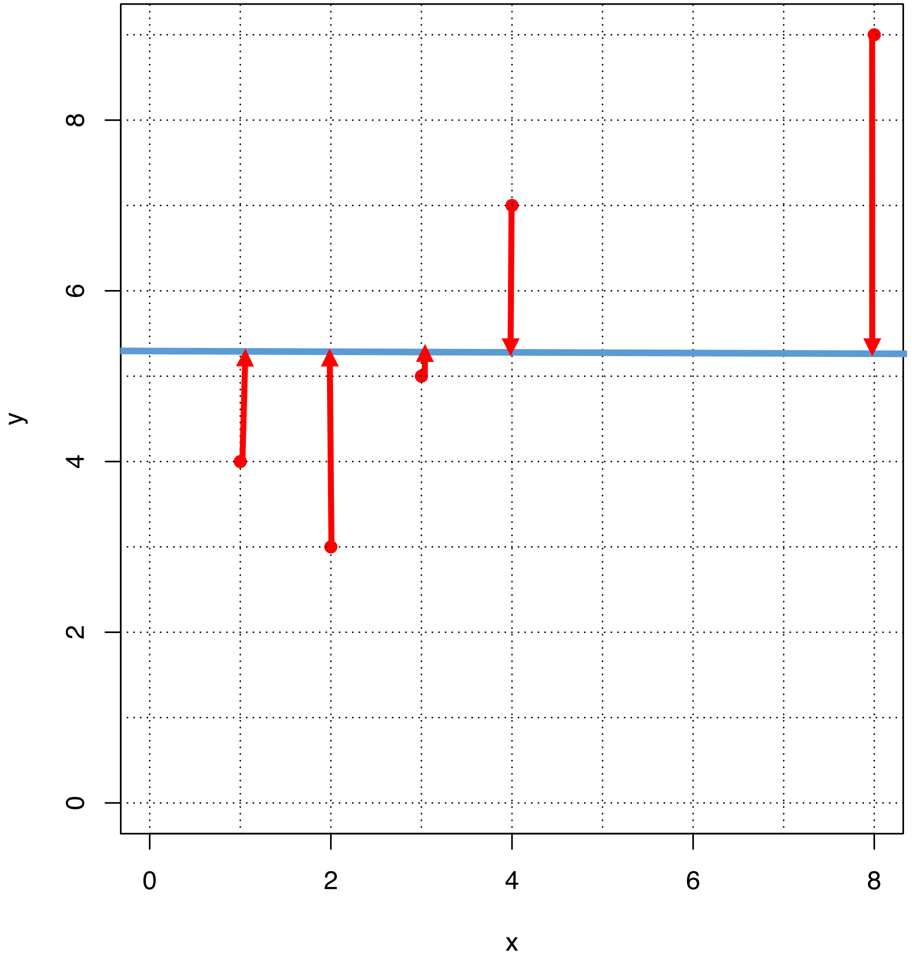
No more guesstimates…



Not a good method!



No more guesstimates…



ε

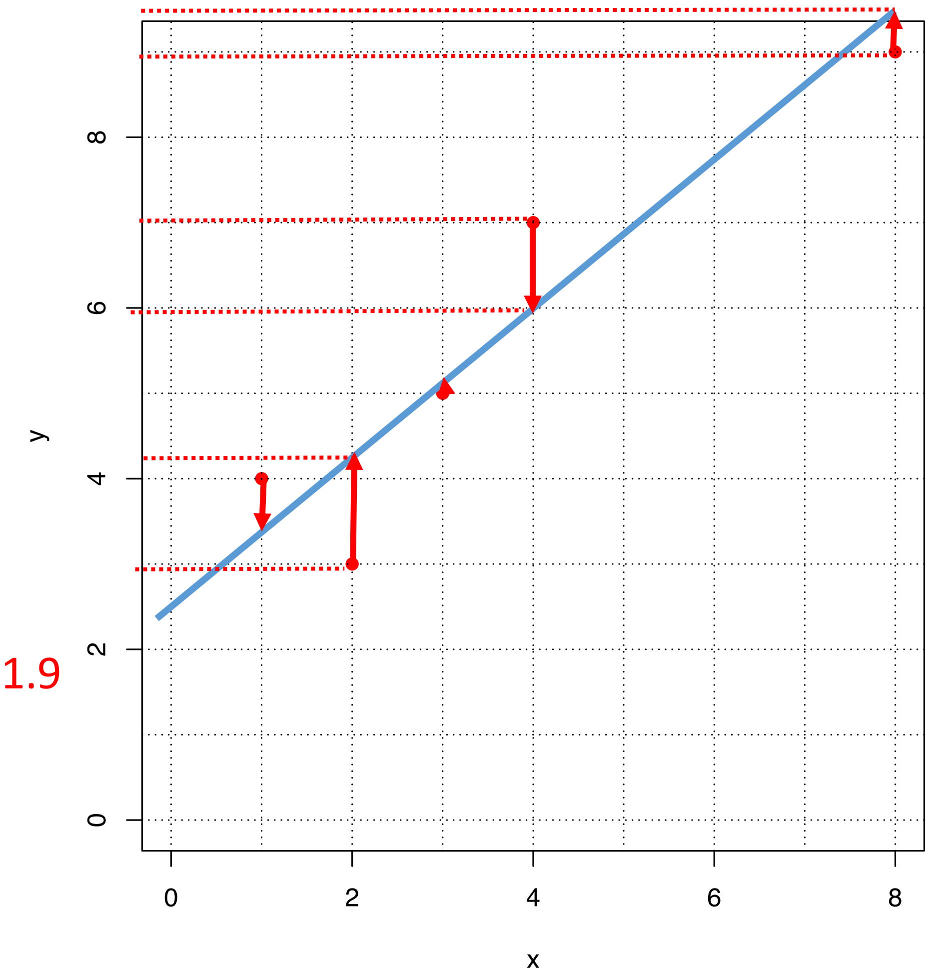
1.2
2.2
0.2
-1.8
-3.4

Sum: -1.6

ε

-0.7
1.2
0.1
1
0.3

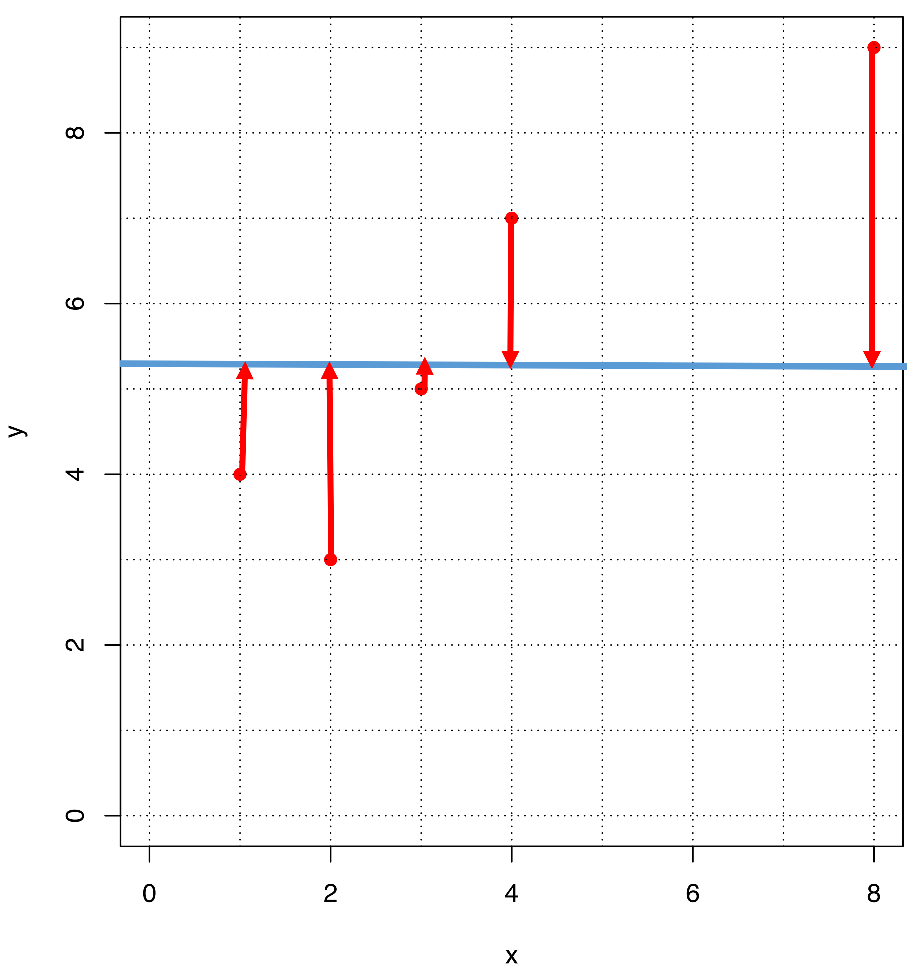
Sum: 1.9



Need to ensure residuals are evaluated as absolute value

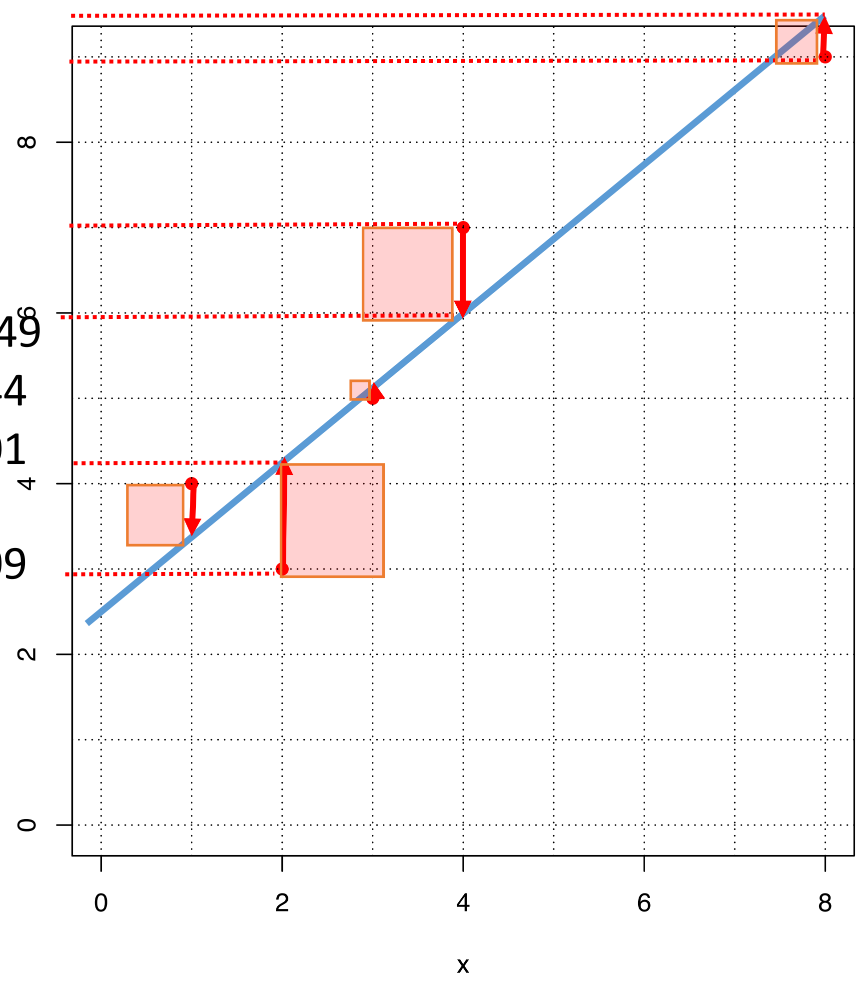
Square all of the residuals!

Sums of squares:

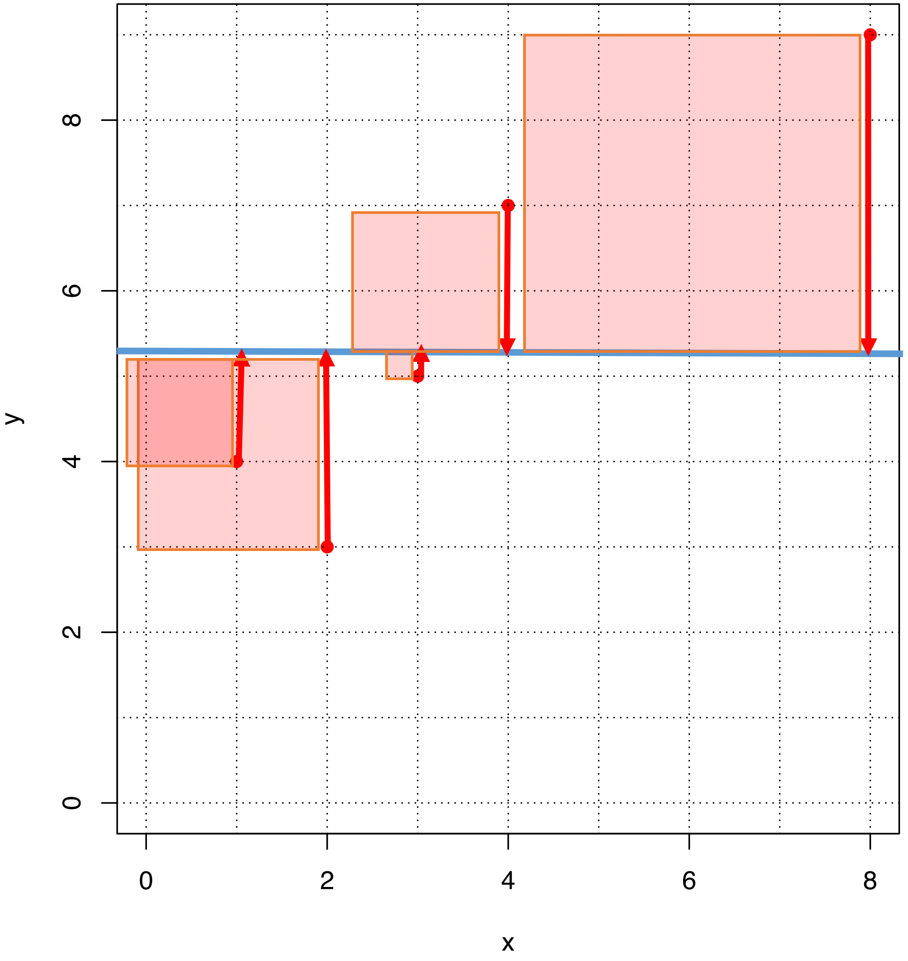


ε
1.2
2.2
0.2
-1.8
-3.4

ε
-0.7 -> 0.49
1.2 -> 1.44
0.1 -> 0.01
1 -> 1
0.3 -> 0.09



Sums of squares:

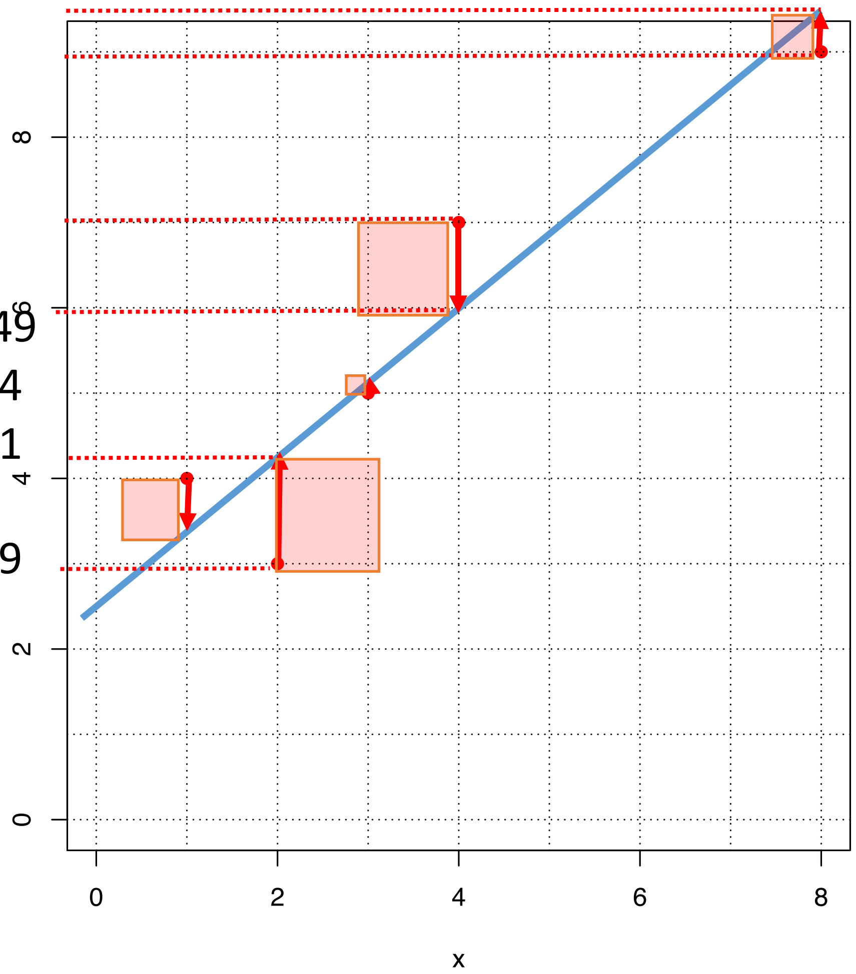


ε

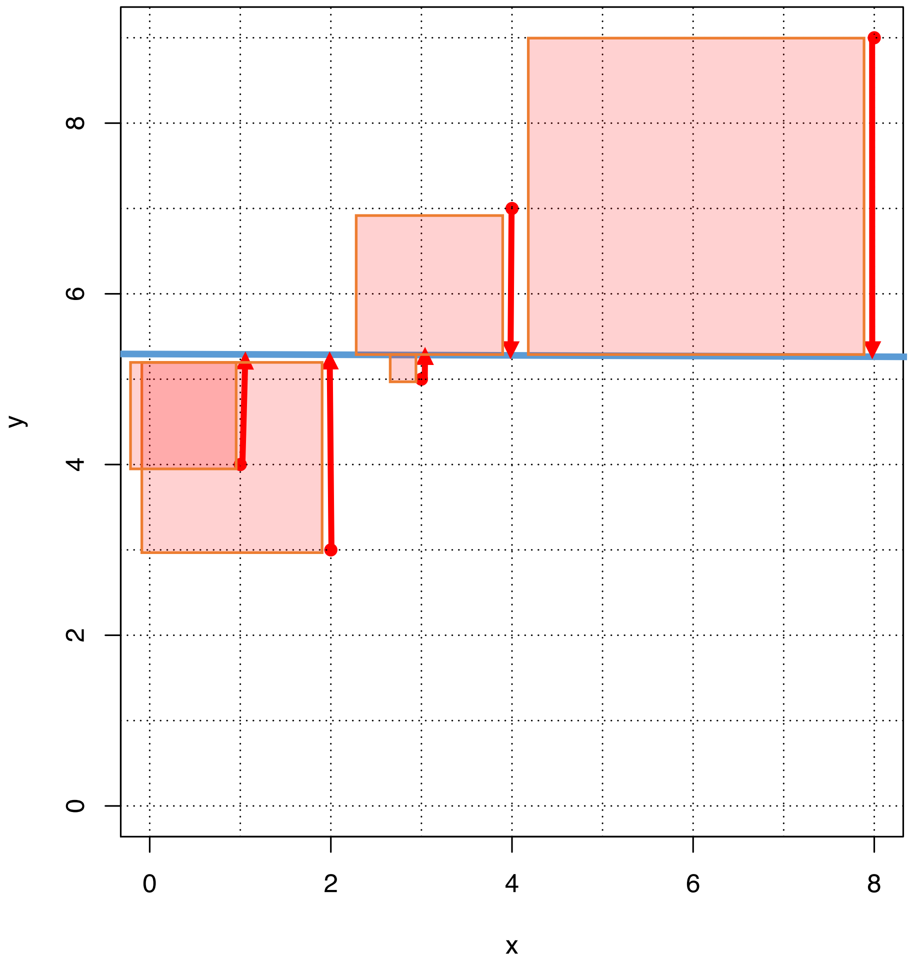
1.2 -> 1.44
2.2 -> 4.84
0.2 -> 0.04
-1.8 -> 3.24
-3.4 -> 11.56

ε

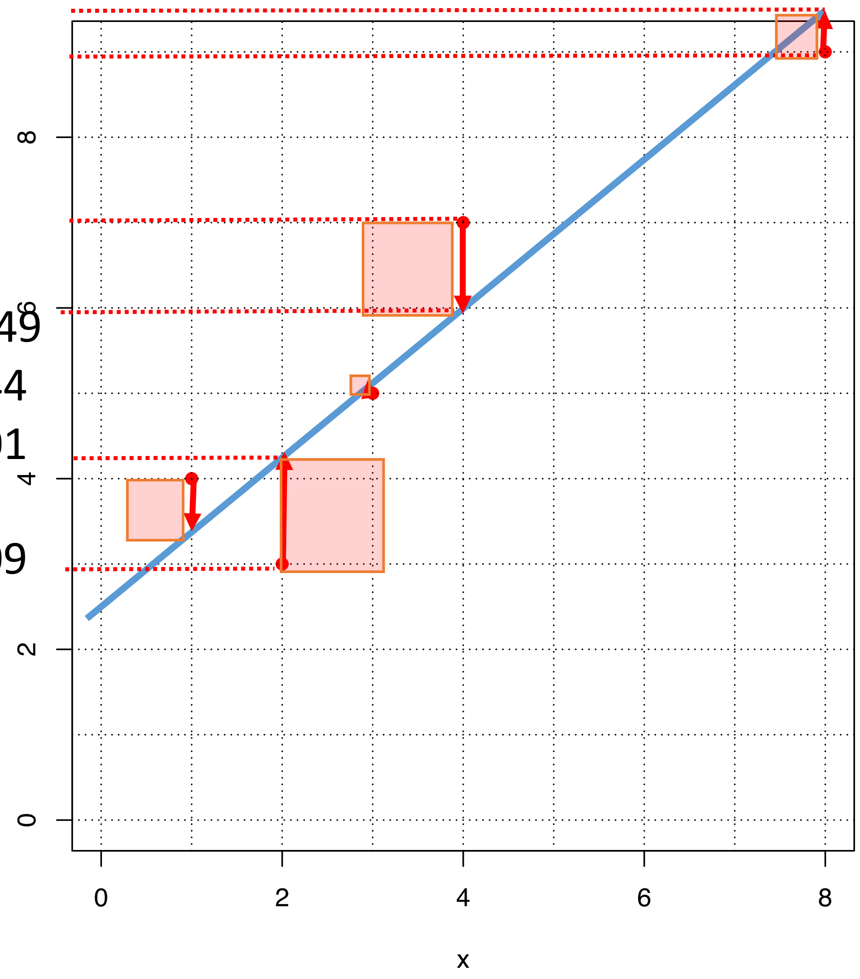
-0.7 -> 0.49
1.2 -> 1.44
0.1 -> 0.01
1 -> 1
0.3 -> 0.09



Sums of squares:



ε	ε
1.2 -> 1.44	-0.7 -> 0.49
2.2 -> 4.84	1.2 -> 1.44
0.2 -> 0.04	0.1 -> 0.01
-1.8 -> 3.24	1 -> 1
-3.4 -> 11.56	0.3 -> 0.09
SS: 21.12	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 ε_i 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 ε_i for the data y_i and x_i

$$find\ minimum\ Q(b0, b1), for\ Q(b0, b1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i\ )^2$$

- 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 ε_i for the data y_i and x_i

$$find\ minimum\ Q(b0, b1), for\ Q(b0, b1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i\ )^2$$

- 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 ε_i for the data y_i and x_i

$$find\ minimum\ Q(b0, b1), for\ Q(b0, b1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

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 ε_i for the data y_i and x_i

$$find\ minimum\ Q(b0, b1),\ for\ Q(b0, b1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

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 ε_i for the data y_i and x_i

$$find\ minimum\ Q(b0, b1),\ for\ Q(b0, b1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

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 ε_i for the data y_i and x_i

$$find\ minimum\ Q(b0, b1), for\ Q(b0, b1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

ALGEBRA

- Solve for b_1, and b_0

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

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 ε_i for the data y_i and x_i

$$find\ minimum\ Q(b0, b1), for\ Q(b0, b1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

ALGEBRA

- Solve for b_1, and b_0

Covariance between x and y

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

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 ε_i for the data y_i and x_i

$$find\ minimum\ Q(b0, b1), for\ Q(b0, b1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

ALGEBRA

- Solve for b_1, and b_0

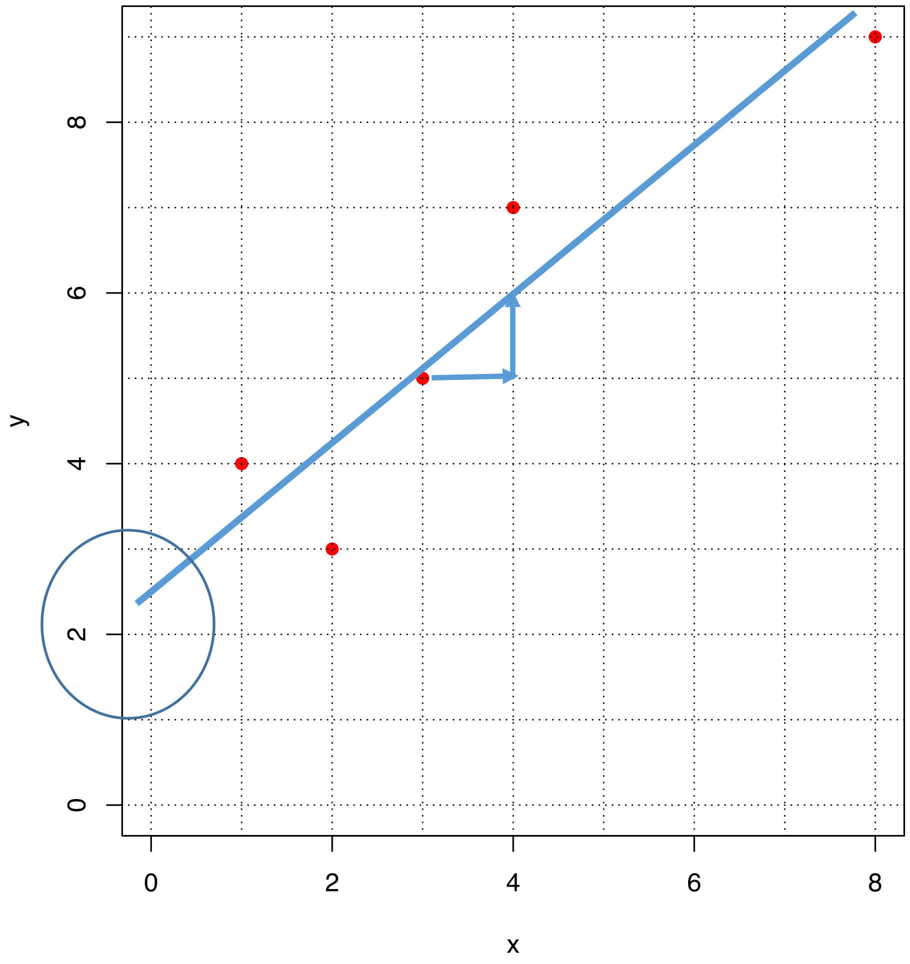
Covariance between x and y

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

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 ε_i?



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

Let’s give this a try:

ε

-0.7 -> 0.49

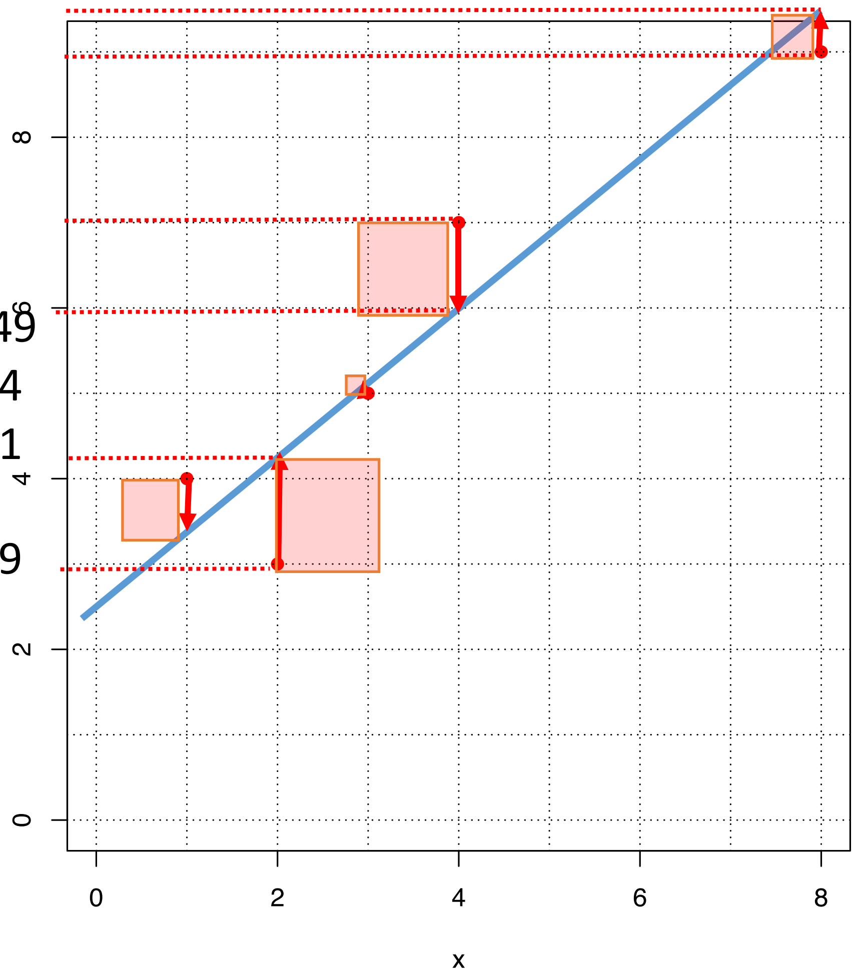
1.2 -> 1.44

0.1 -> 0.01

1 -> 1

0.3 -> 0.09

SS: 3.03



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

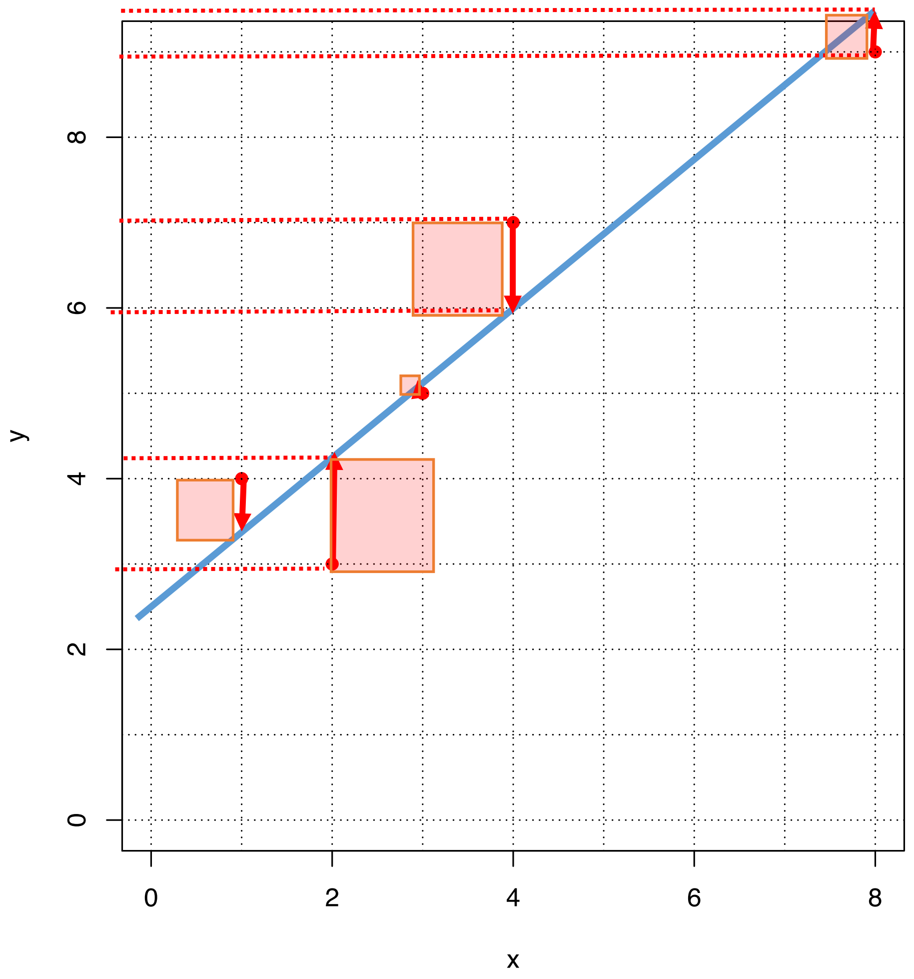
SS: 3.03

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

$$b_1 = \frac{Cov\,[x,y]}{\sigma_x^2}$$

Let's give this a try:

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



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

SS: 3.03

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

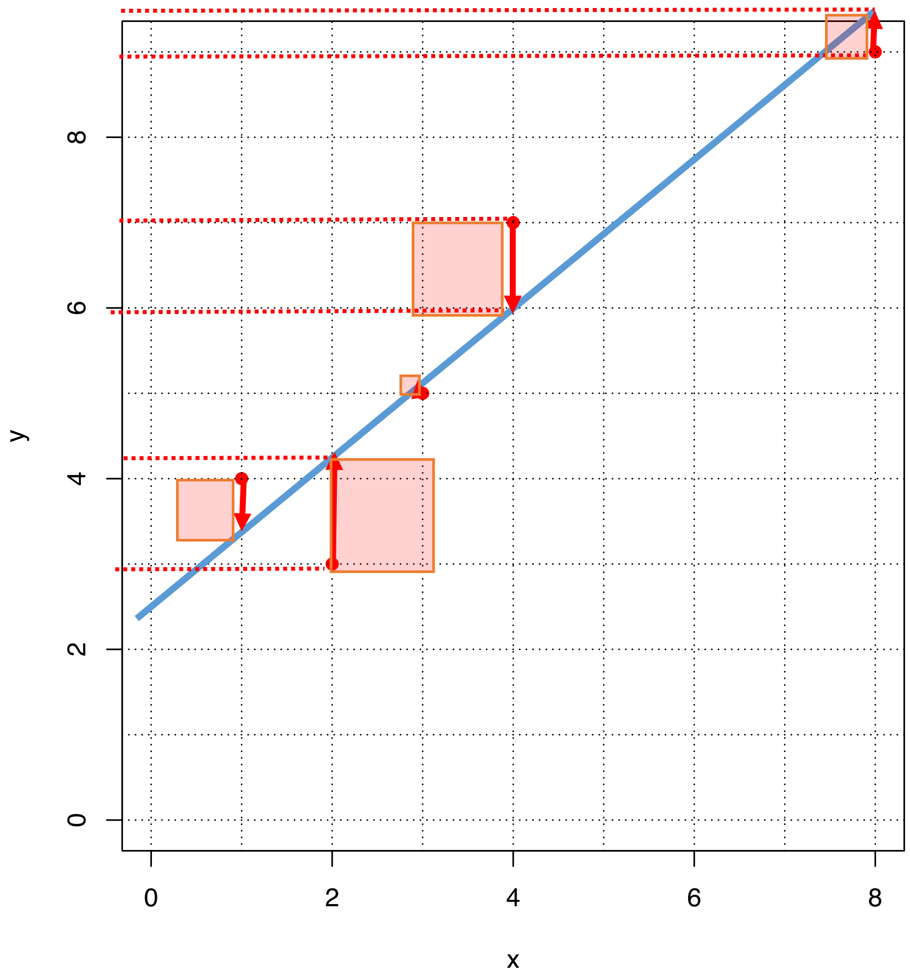
$$b_1 = \frac{Cov\,[x,y]}{\sigma_x^2}$$

Let’s give this a try:

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

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

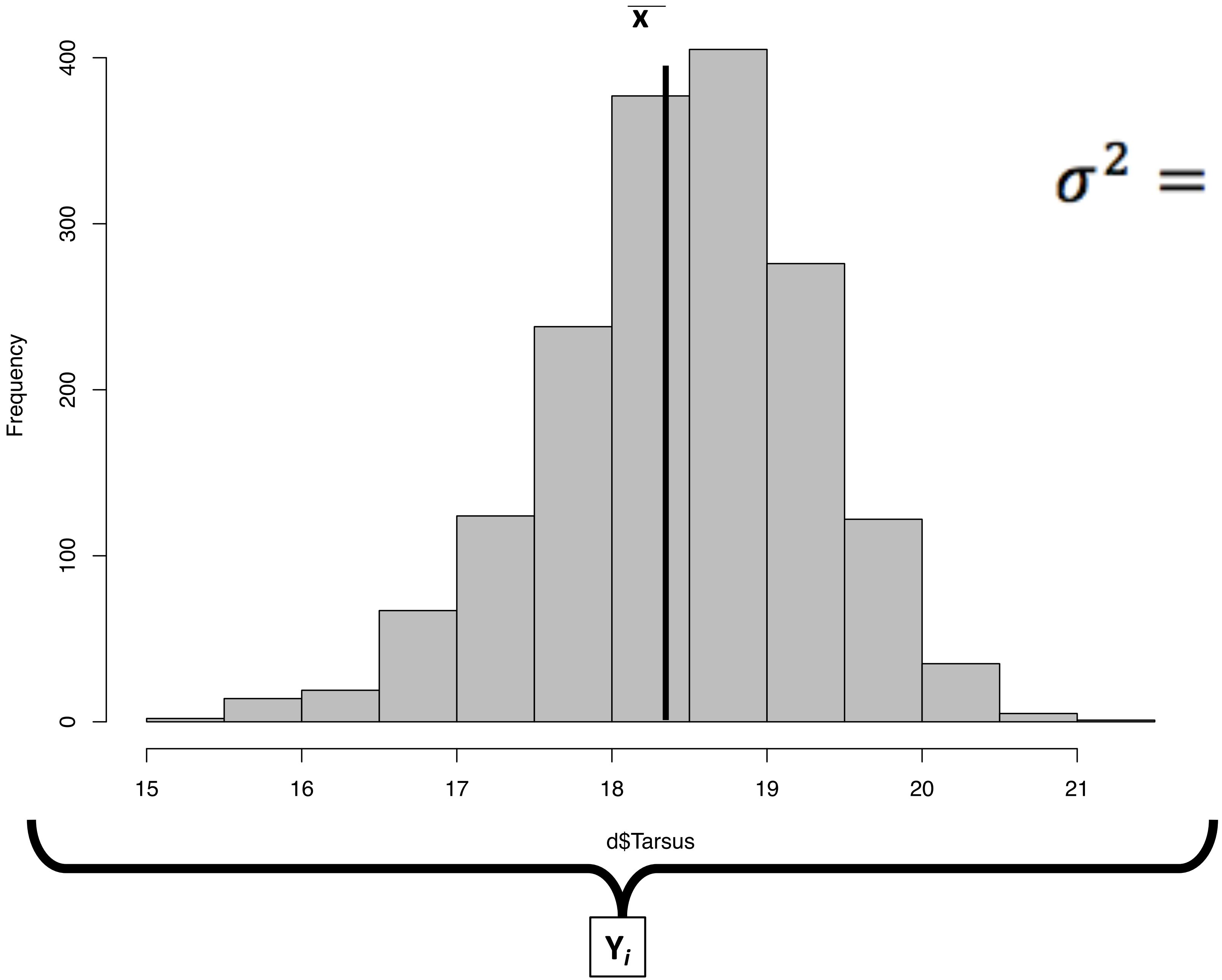
$$b_1 = \frac{Cov\,[x,y]}{\sigma_x^2}$$



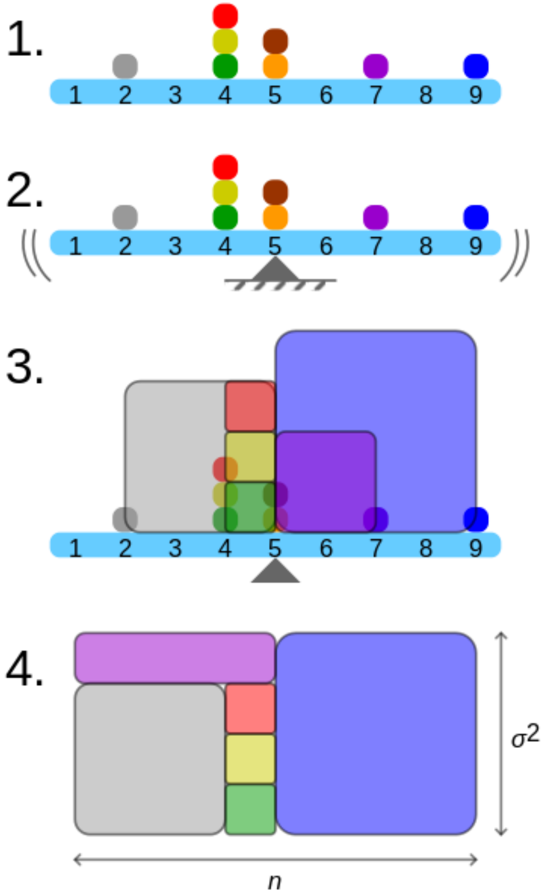
$$y_i = 2.2 + 1x_i + \varepsilon_i$$

SS: 3.03

VARIANCE



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



Let’s give this a try:

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

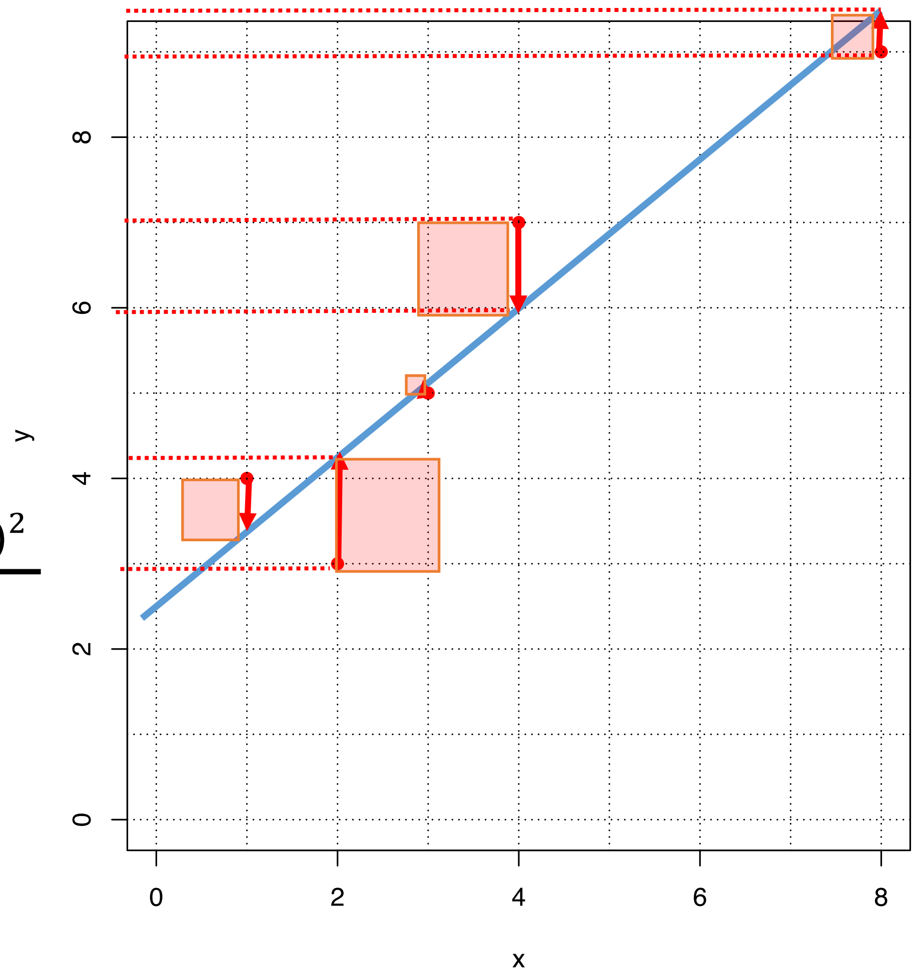
$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

$$= \frac{(1-3.6)^2 + (2-3.6)^2 + (3-3.6)^2 + (4-3.6)^2 + (8-3.6)^2}{4}$$

$$= 7.3$$

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

$$b_1 = \frac{Cov\,[x,y]}{\sigma_x^2}$$



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

SS: 3.03

Let's give this a try:

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

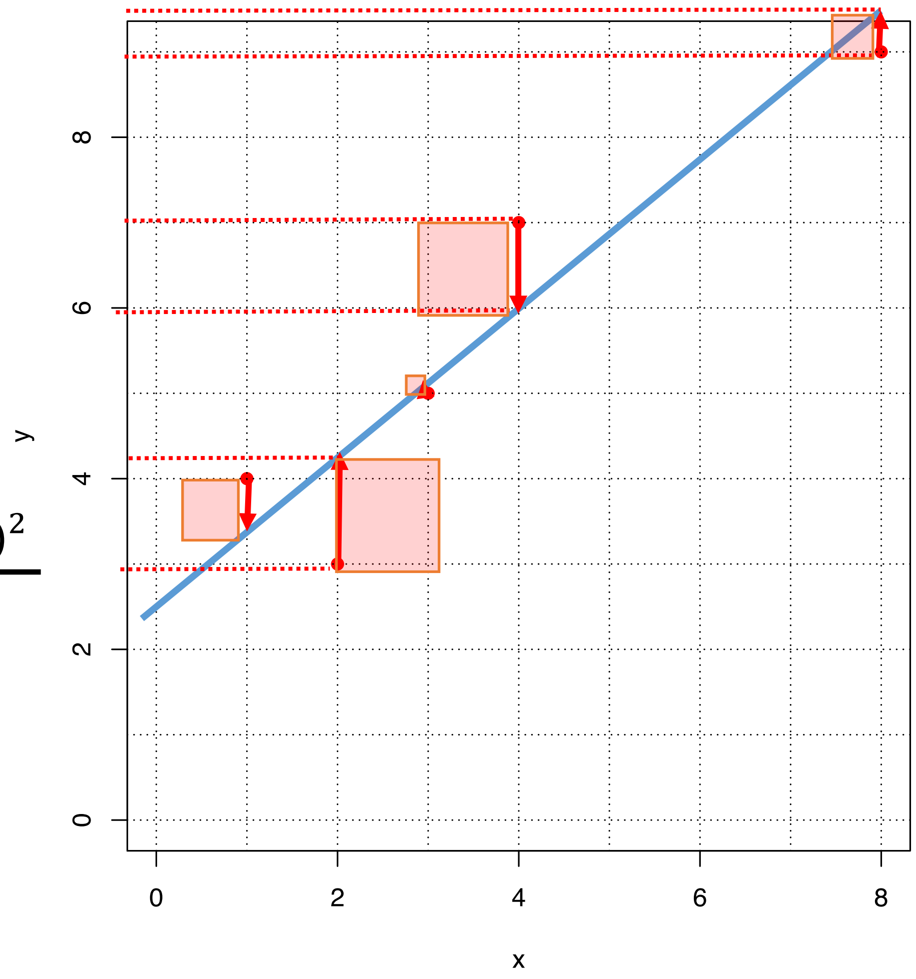
$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

$$= \frac{(1-3.6)^2 + (2-3.6)^2 + (3-3.6)^2 + (4-3.6)^2 + (8-3.6)^2}{4}$$

$$= 7.3$$

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

$$b_1 = \frac{Cov\,[x,y]}{\sigma_x^2} = \frac{Cov\,[x,y]}{7.3}$$



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

SS: 3.03

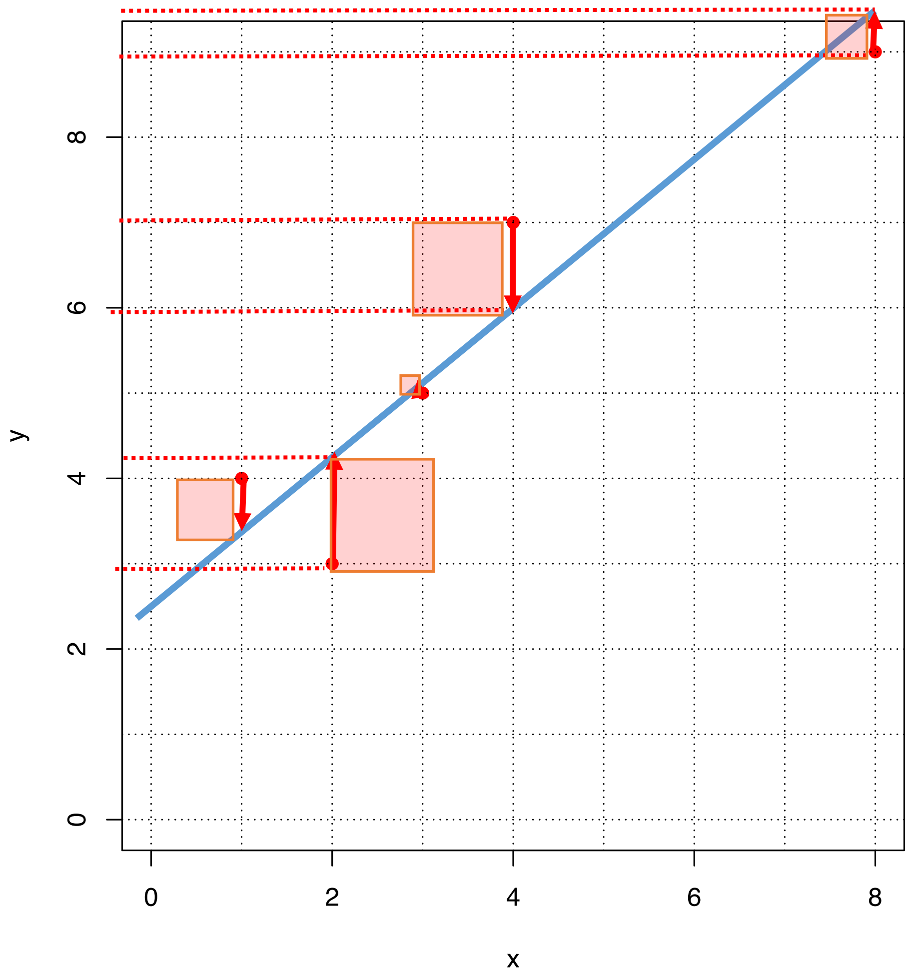
Let's give this a try:

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

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

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

$$b_1 = \frac{Cov\,[x,y]}{\sigma_x^2} = \frac{Cov\,[x,y]}{7.3}$$



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

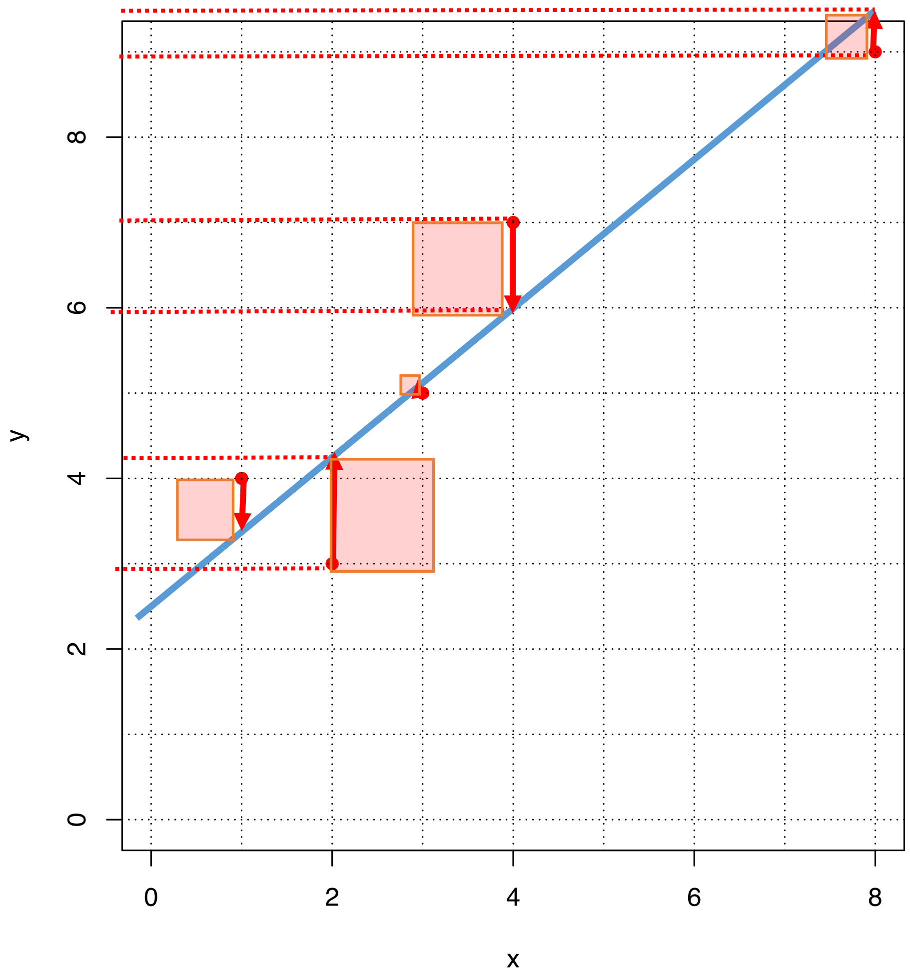
SS: 3.03

Let’s give this a try:

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

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

$= 6.05$



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

SS: 3.03

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

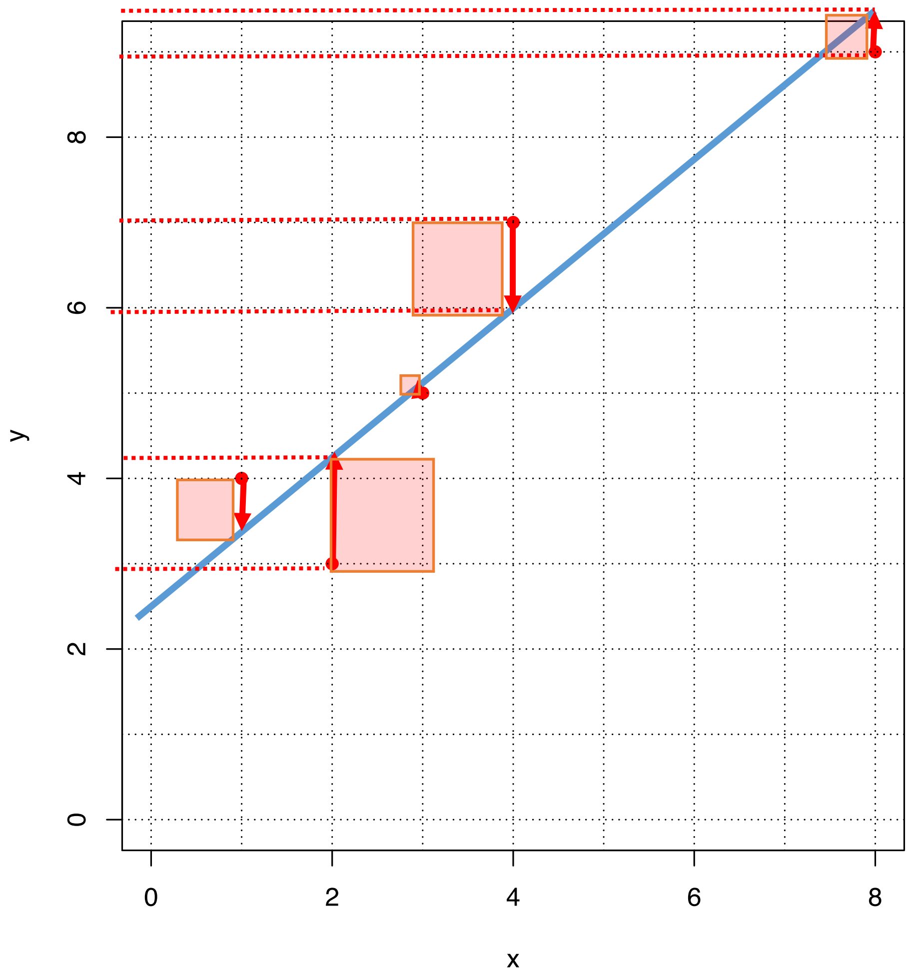
$$b_1 = \frac{Cov\ [x,y]}{\sigma_x^2} = \frac{Cov\ [x,y]}{\mathbf{7.3}}$$

Let’s give this a try:

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

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

$= 6.05$



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

SS: 3.03

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

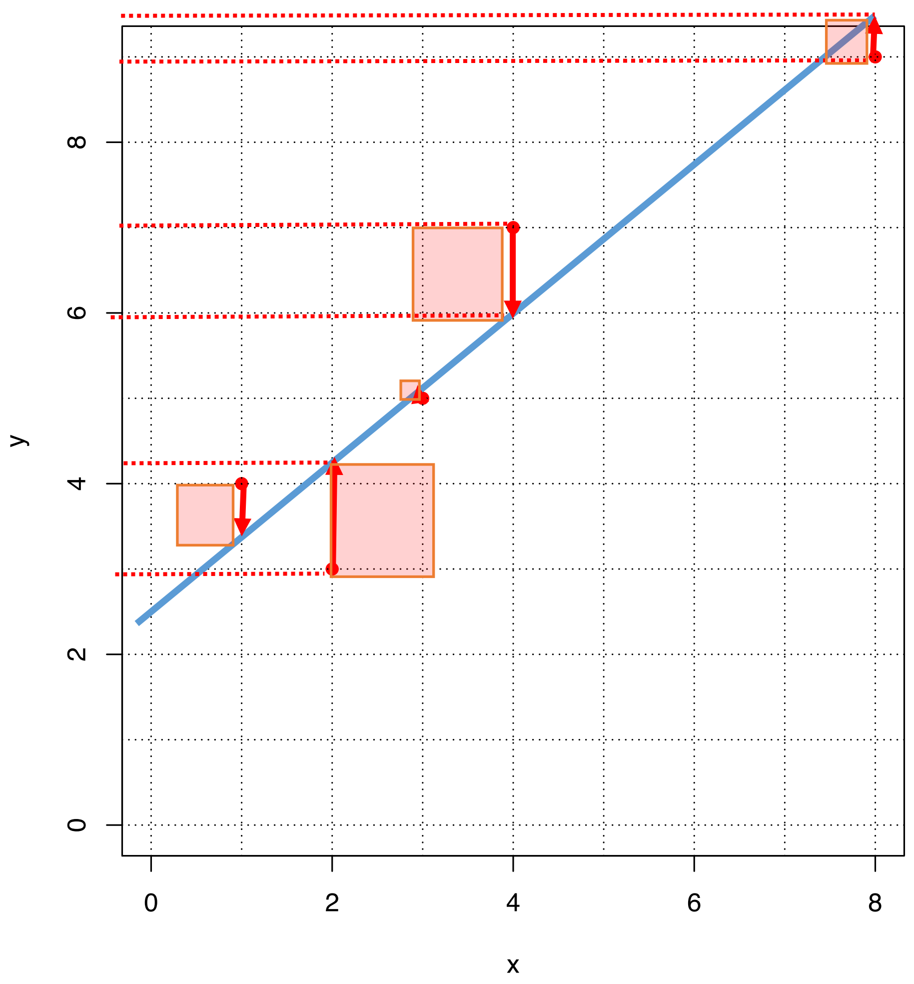
$$b_1 = \frac{Cov\,[x,y]}{\sigma_x^2} = \frac{6.05}{7.3}$$

Let's give this a try:

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

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

$= 6.05$



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

SS: 3.03

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

$$b_1 = \frac{Cov\,[x,y]}{\sigma_x^2} = \frac{\mathbf{6.05}}{\mathbf{7.3}} = 0.83$$

Let's give this a try:

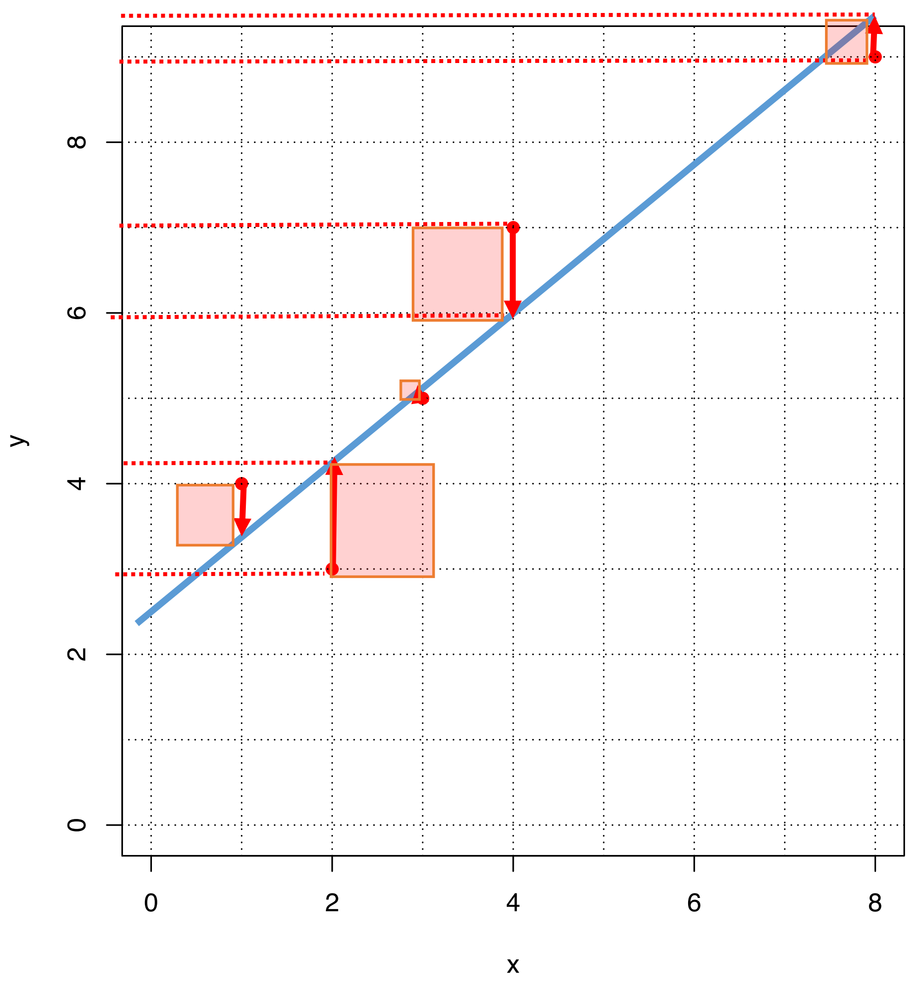
x, y
1,4
2,3
3.5
4,7
8,9

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

$= 6.05$

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

$$b_1 = \frac{Cov\ [x,y]}{\sigma_x^2} = \frac{\mathbf{6.05}}{\mathbf{7.3}} = 0.83$$



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

SS: 3.03

Let's give this a try:

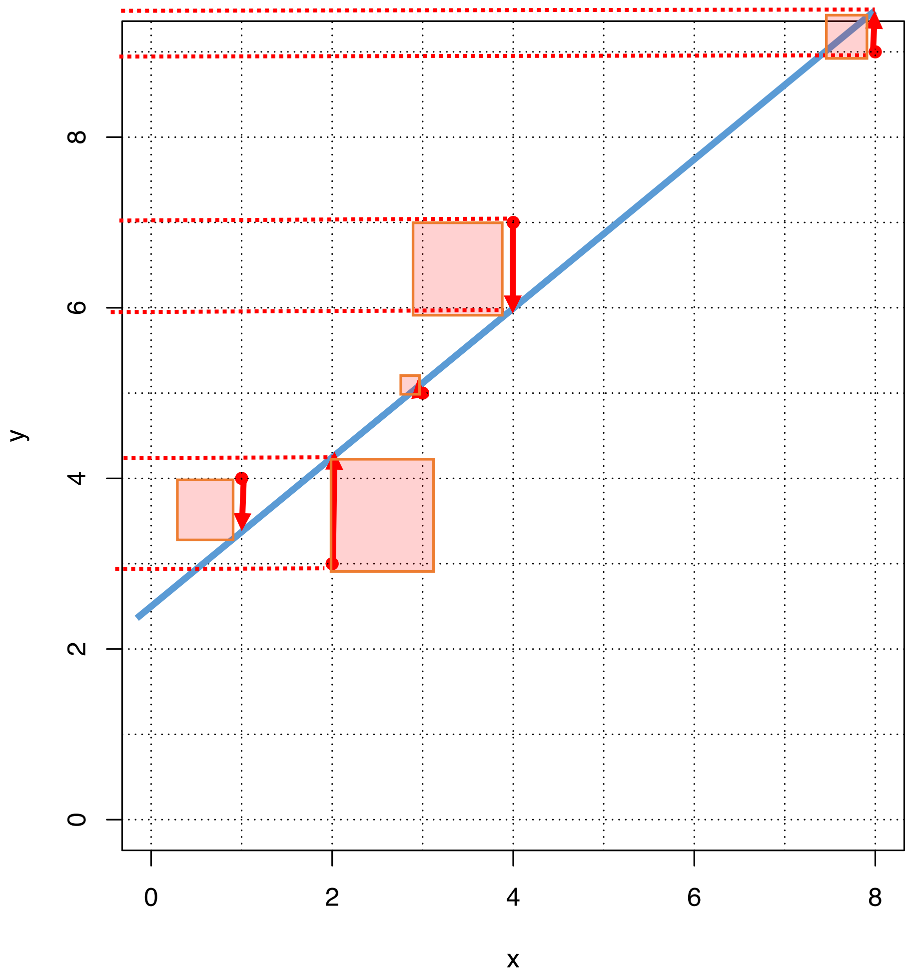
x, y
1,4
2,3
3.5
4,7
8,9

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

$= 6.05$

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

$= 5.6 – 0.83 * 3.6$



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

SS: 3.03

$$b_1 = \frac{Cov\ [x,y]}{\sigma_x^2} = \frac{\mathbf{6.05}}{\mathbf{7.3}} = 0.83$$

Let's give this a try:

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

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

= 6.05

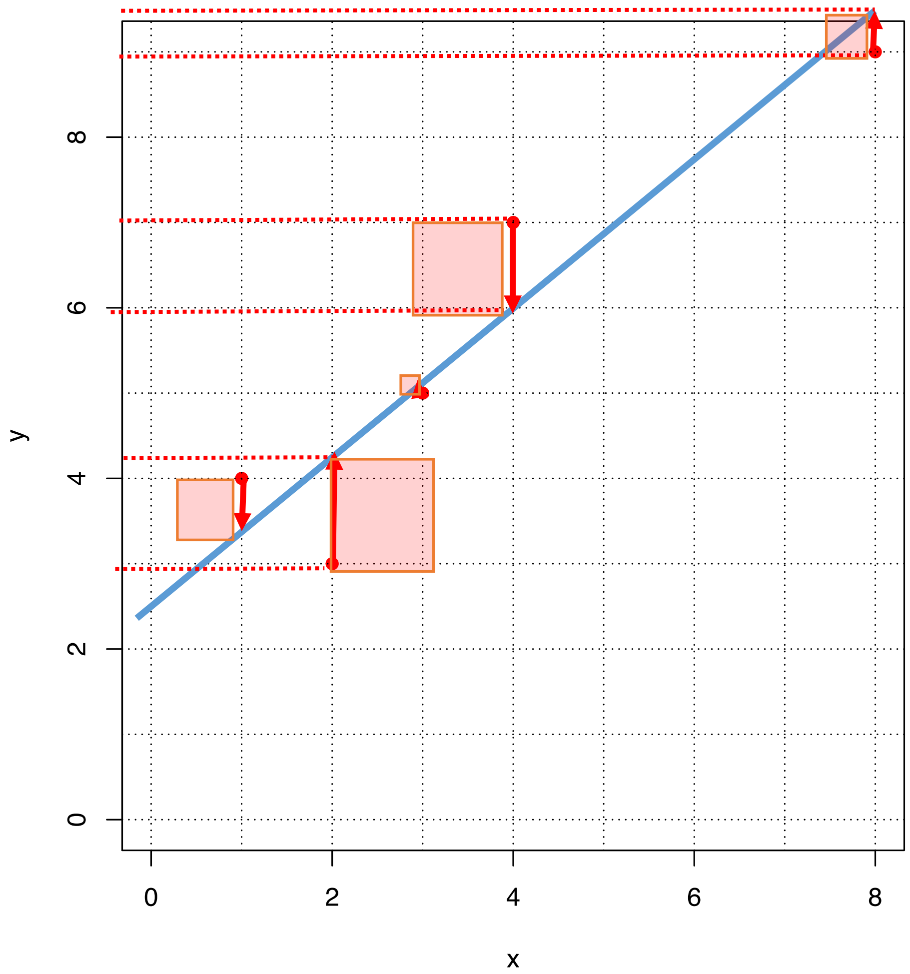
$$b_0 = \bar{y} - b_1\bar{x}$$

= 5.6 – 0.83 * 3.6

= 5.6 – 2.99

= 2.5

$$b_1 = \frac{Cov\ [x,y]}{\sigma_x^2} = \frac{\mathbf{6.05}}{\mathbf{7.3}} = 0.83$$



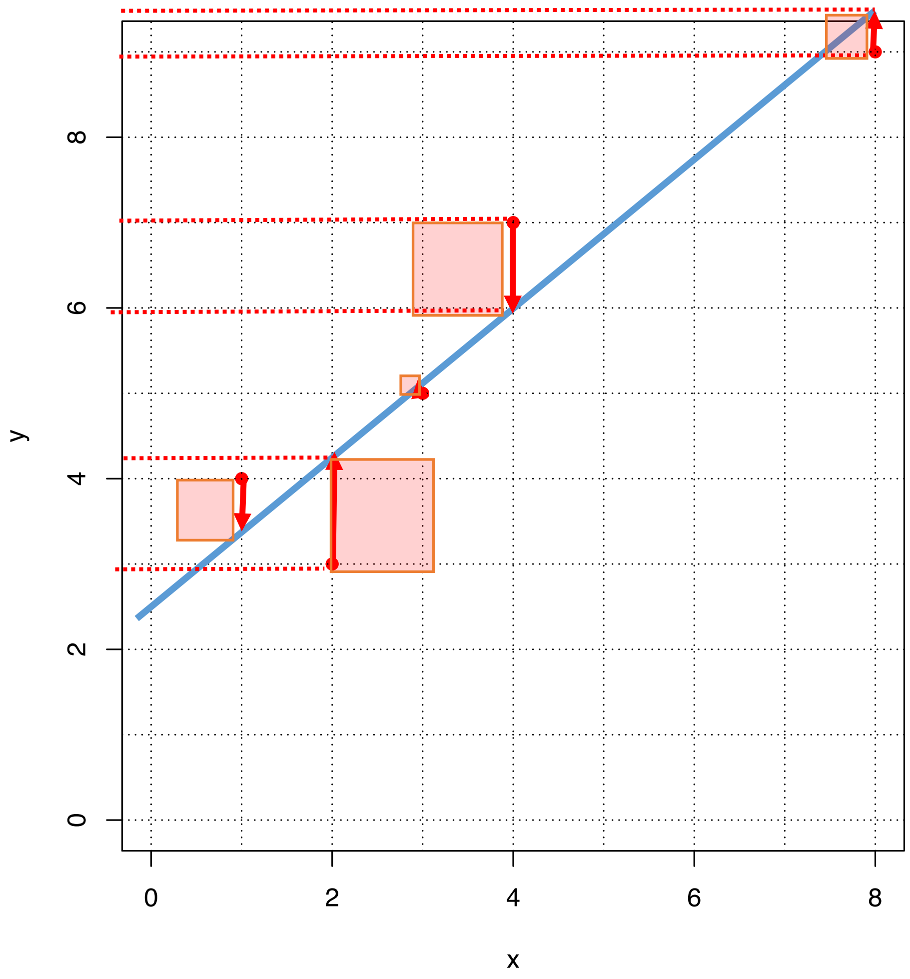
$$y_i = 2.2 + 0.83x_i + \varepsilon_i$$

SS: 3.03

R^2

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

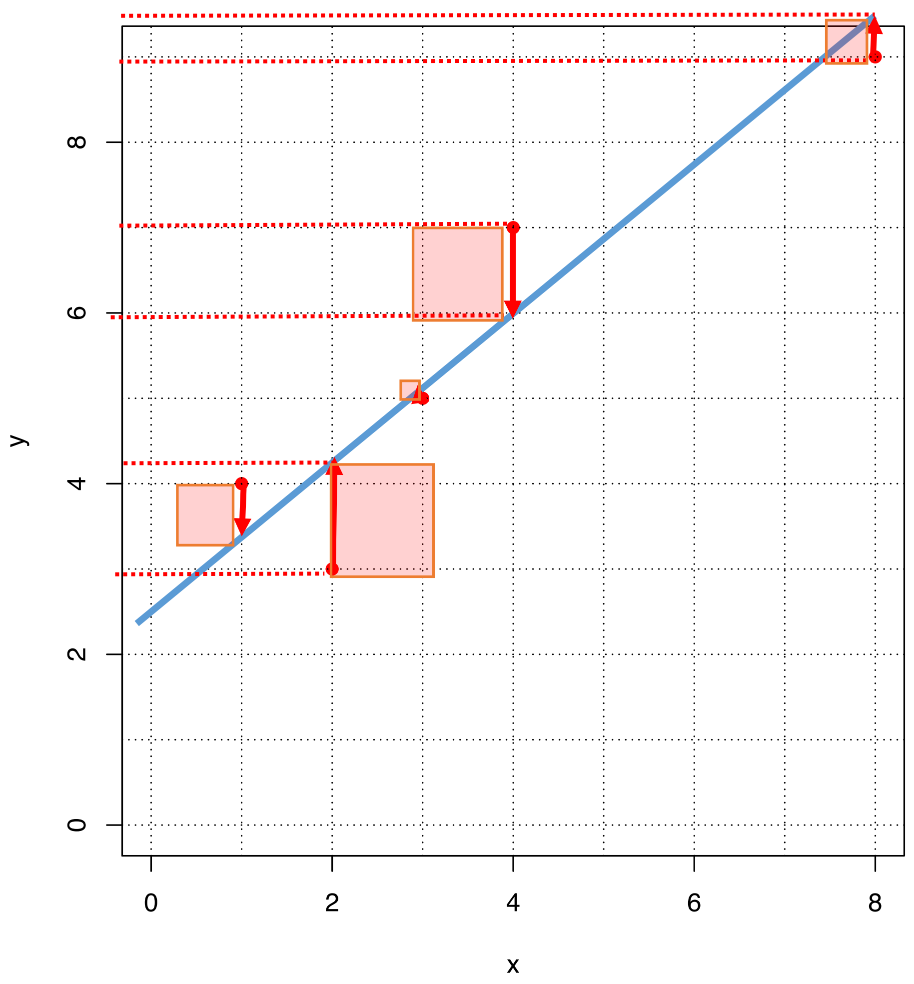
$$R^2 = 1 - \frac{SS_{residuals}}{SS_{total}}$$



R^2

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

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

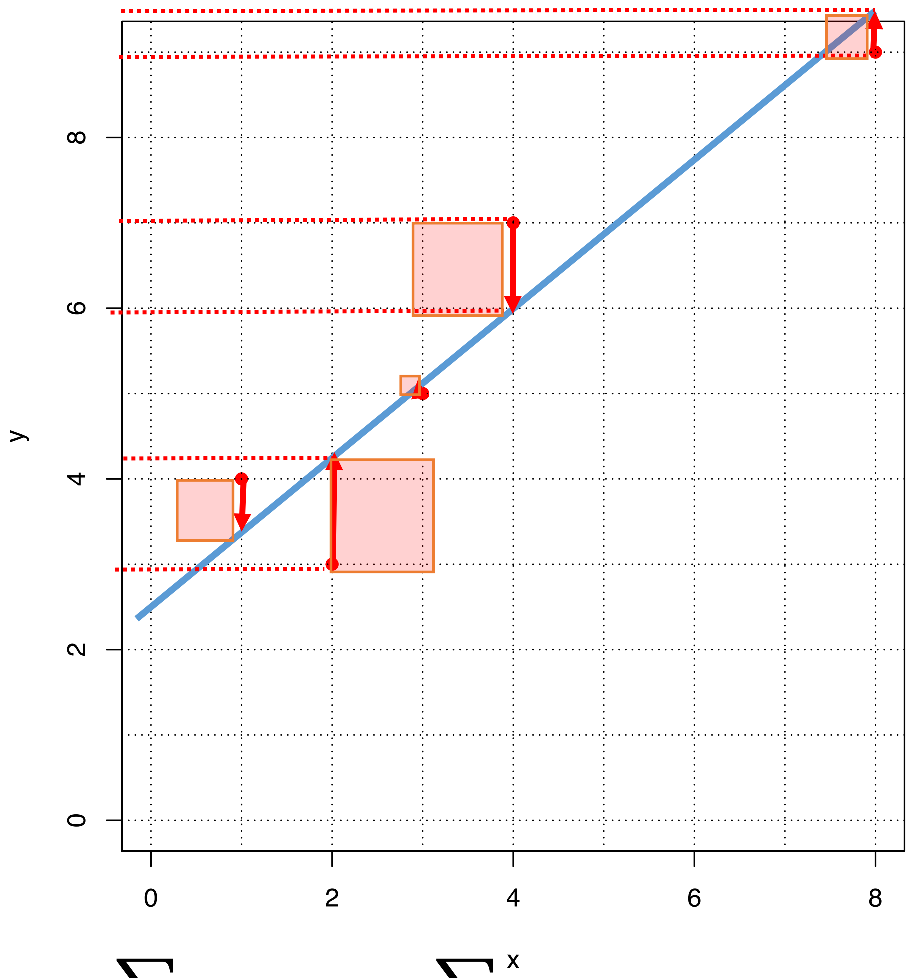


Wait, what?

R^2

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

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

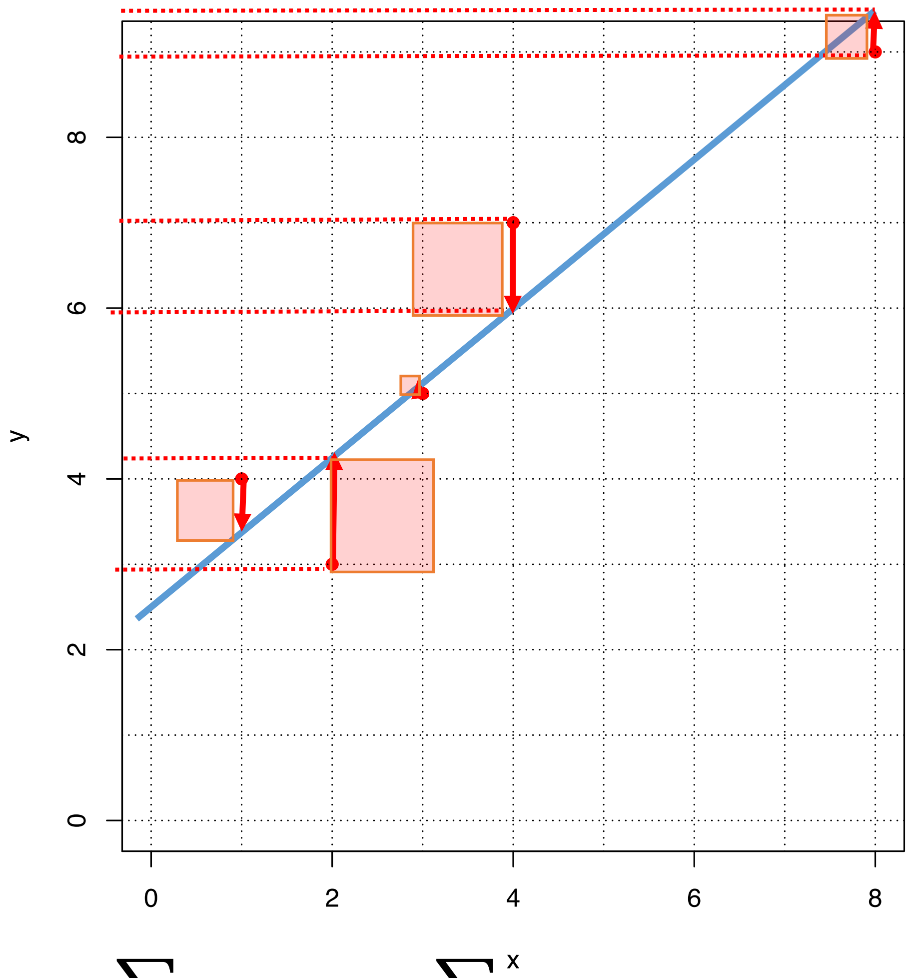


We know what SS_{res} is – the residual sum of squares. $\sum (y_i - x_i)^2 = \sum (\varepsilon_i)^2$

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

$$\sum (y_i - x_i)^2 = \sum (\varepsilon_i)^2$$

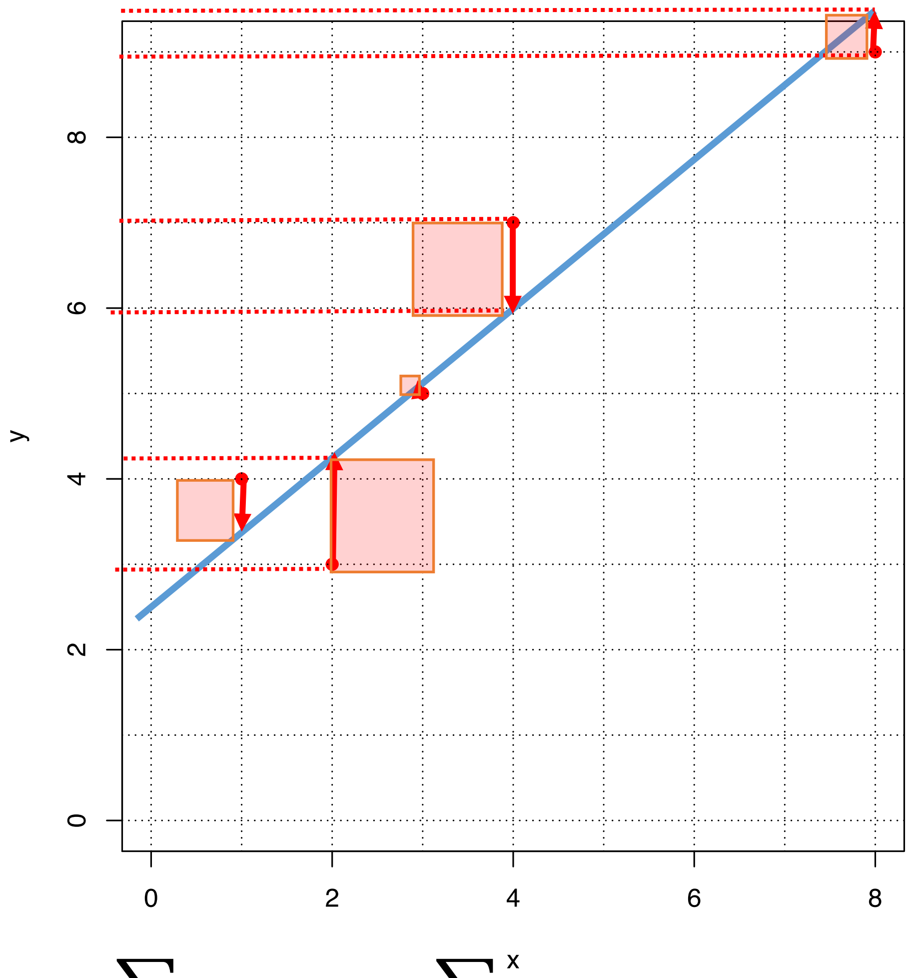
The total sum of squares is this:

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

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

$$\sum(y_i - x_i)^2 = \sum(\varepsilon_i)^2$$

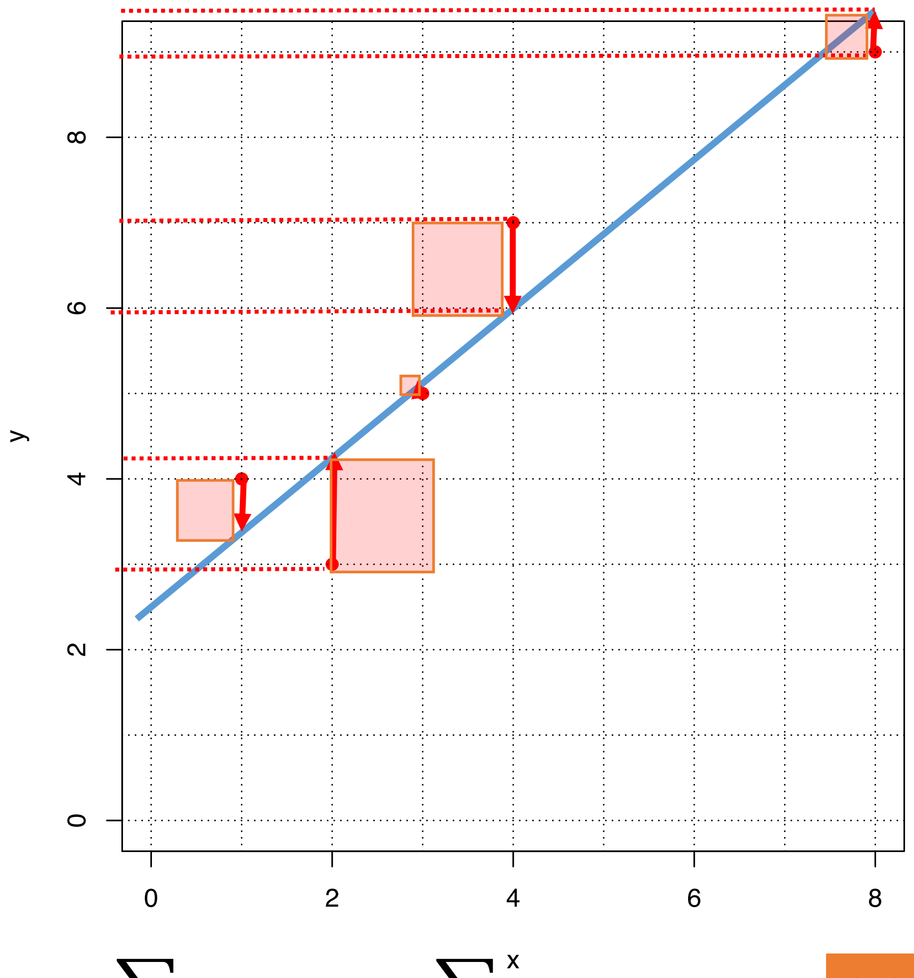
The total sum of squares is this:

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

R^2

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

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



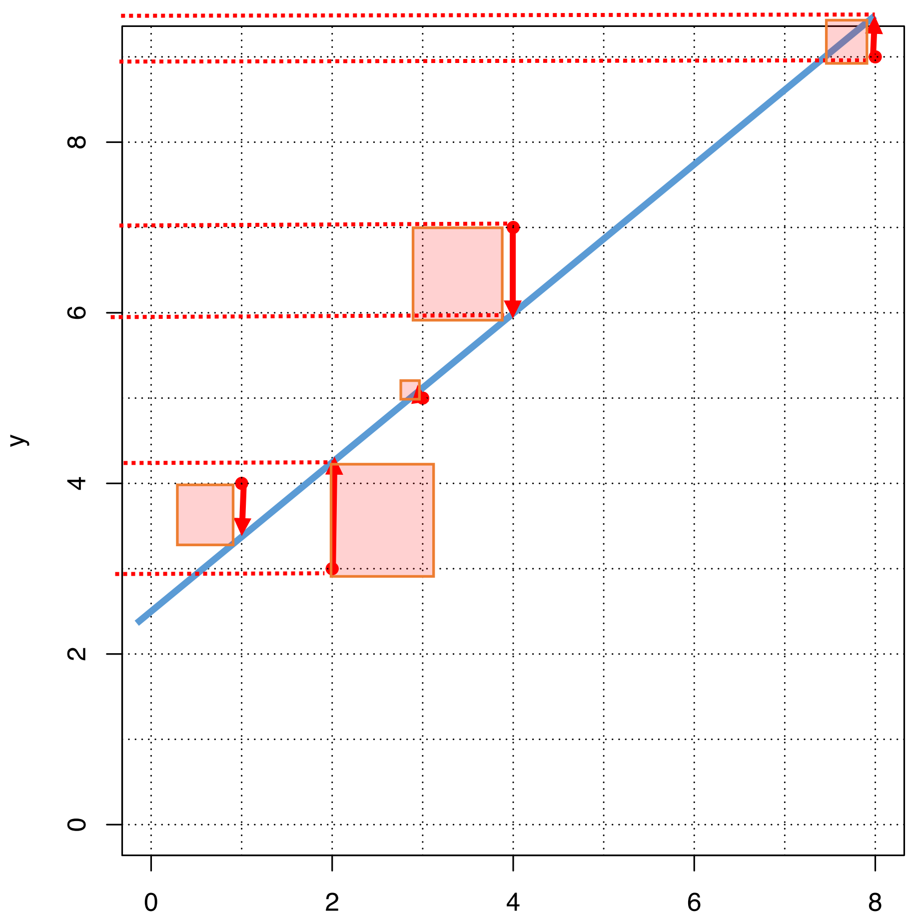
x, y
1,4
2,3
3.5
4,7
8,9

We know what SS_{res} is – the residual sum of squares $\sum (y_i - x_i)^2 = \sum (\varepsilon_i)^2$ 3.03

The total sum of squares is this: $\sum (y_i - \bar{y})^2 = \sigma^2 * (n-1)$ 5.8 * 4
23.2

R^2

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



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

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

= 1 – 3.03/23.2
= 0.87

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

$$\sum (y_i - x_i)^2 = \sum^{x} (\varepsilon_i)^2$$

3.03

The total sum of squares is this:

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

5.8 * 4
23.2

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

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

Exercise – no hand-out

$y_i = 2.2 + 0.83x_i + \varepsilon_i$

ε
-0.7 -> 0.49
1.2 -> 1.44
0.1 -> 0.01
1 -> 1
0.3 -> 0.09

SS: 3.03

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

```
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

Questions:
What is confirmed?
What is different? Why?