Questions about the class…

- Registration/waitlist questions:
  - gradinfo@stat.ubc.ca
- Labs start this Thursday.
- Lab registration (about 10 students):
  - Send an email with what labs are available to you:
  gradinfo@stat.ubc.ca
- Midterm; Lecture Slides; Webwork
  - webwork.elearning.ubc.ca

Questions about the class…

- Midterm; Lecture Slides; Webwork…
- Pre-requisite knowledge:
  - One sample and two sample t-tests
  - Hypothesis tests
  - Confidence Intervals
  - Type 1 error and Type 2 error
  - probability density function (pdf), cumulative density function (cdf).
  - Properties of a Normal distribuiton.

Stat 306: Finding Relationships in Data.

Lecture 2

Least Squares for one predictor

t-test

Age vs. Money

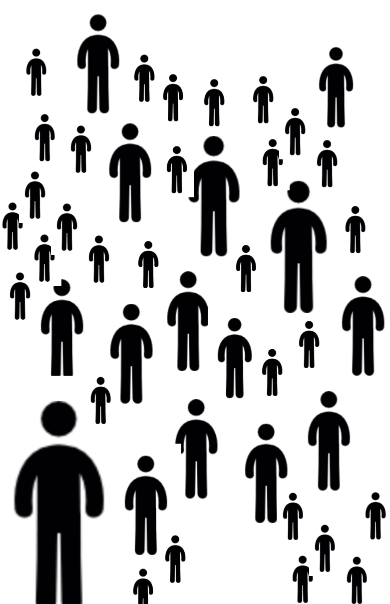
Dependent variable

X → old (0), young (1)

Independent variable

Y → dollars ($) In bank account

Population



Population parameters

$$\mu_0 , \mu_1 , \sigma^2$$

Hypothesis Test

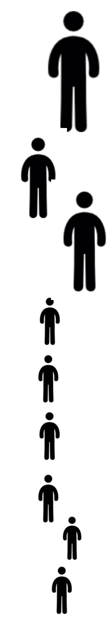
$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 \neq \mu_1$$

Sample statistics

$$\bar{y}_0 = 56$$
$$\bar{y}_1 = 27$$
$$\bar{y}_0 - \bar{y}_1 = 29$$
$$s_p = 10.81$$

t = 2.68, df = 7
p-value = 0.03
95% C.I. = [3.4, 54.6]

Sample, n=9



x	y
old	71
old	54
old	43
young	45
young	21
young	11
young	30
young	45
young	10

linear regression

Age vs. Money

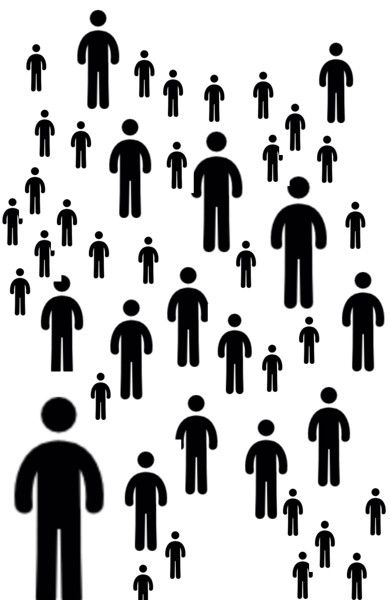
PREDICTOR *variable*

$\chi \longrightarrow$ Age in Years

RESPONSE *variable*

$\Upsilon \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$$\beta_0, \beta_1, \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

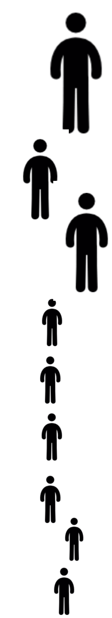
Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For statistic b_1 :

95% C.I. = [0.05, 1.05]
p-value = 0.036

Sample, n=9



χ	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Chapter 2

Simple linear regression

In forming prediction equations, there are typically several explanatory variables $x_1, \ldots, x_p$ that affect the response variable y. Given data vectors $(x_{i1}, \ldots, x_{ip}, y_i)$ for $i = 1, \ldots, n$, the goal is the find a prediction equation that goes through the middle of the data points.

This chapter starts with the study of the simple case of one explanatory variable ($p = 1$). The data vectors are written as (x_i, y_i) for $i = 1, \ldots, n$ and the simplest prediction equation to consider is a straight line that goes through the middle of the scatterplot of y versus x.

The theory of estimation of the prediction equation with least squares involves calculus, and the statistical inference for predictions involve a statistical or probability model for the random deviations from a prediction equation. The usual assumption (for the first attempted prediction equation) is that deviations from a theoretical prediction equation are independent normal random variables with a mean of 0 and a variance that does not depend on the values of the explanatory variables.

For one explanatory variable, derivations of estimators and interval estimates can be done without matrices, and it is easier to show the meaning of different quantities via scatterplots. The same steps with matrices and vectors will be carried out in Chapter 3 when there is more than one explanatory variable.

- We will start with the simple case of one **explanatory variable**.
- The **data vectors** are written as (x_i , y_i) for $i = 1, . . . , n$.
- The simplest **prediction equation** to consider is a straight line that goes through the middle of the scatterplot of y versus x.

Age vs. Money

"explanatory variable"

PREDICTOR *variable*

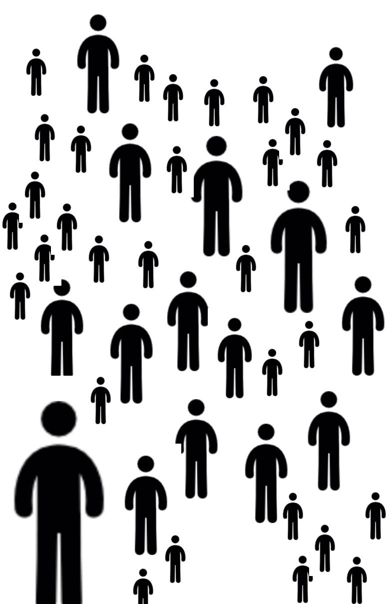
$X \longrightarrow$ Age in Years

(or dependent)

RESPONSE *variable*

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$$\beta_0 , \beta_1 , \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

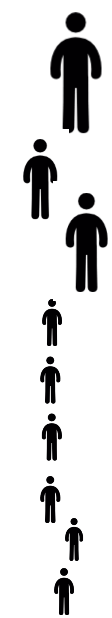
Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

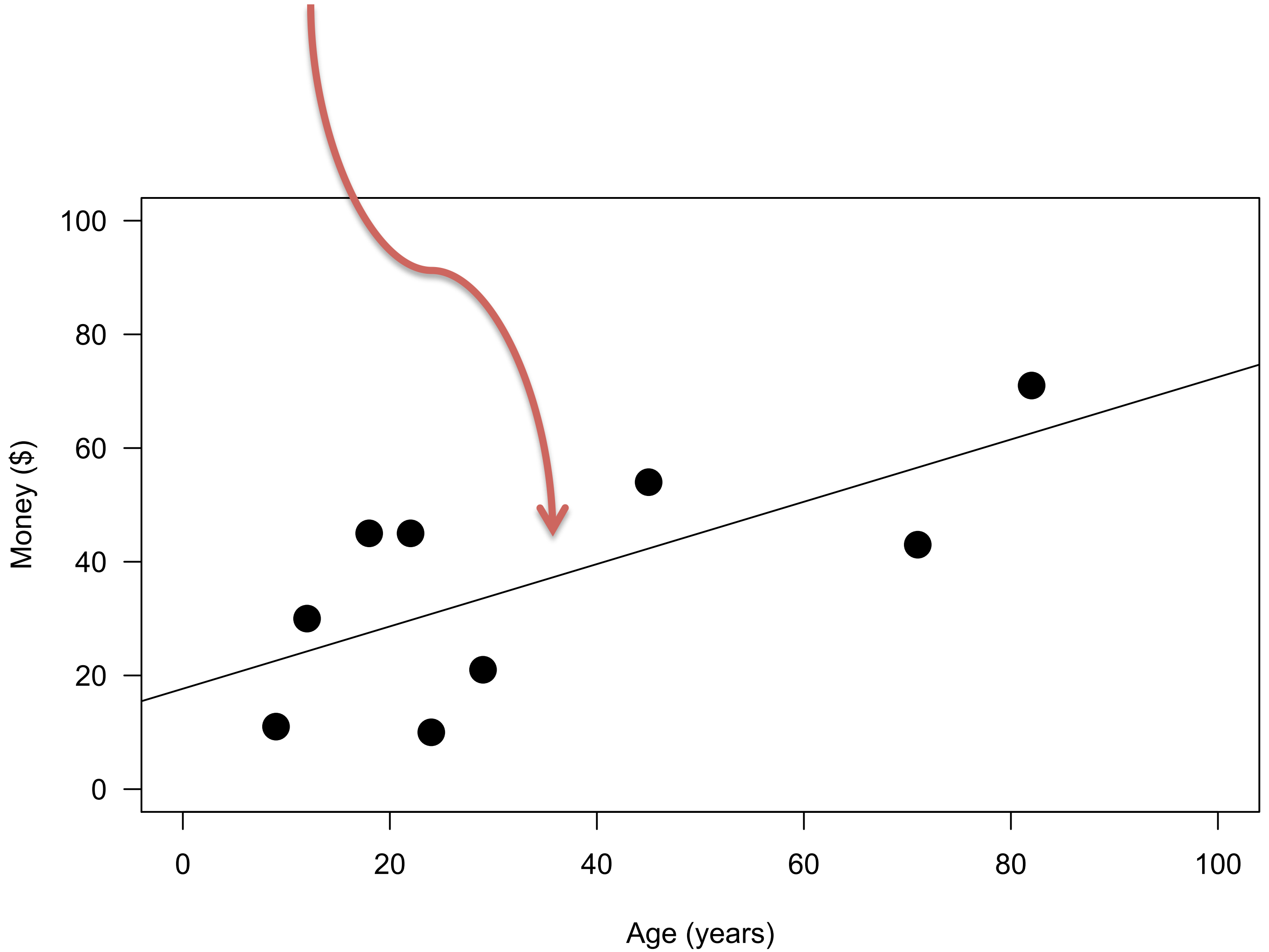
95% C.I. $= [0.05, 1.05]$
p-value $= 0.036$

data vectors



i	x	y
1	82	71
2	45	54
3	71	43
4	22	45
5	29	21
6	9	11
7	12	30
8	18	45
9	24	10

prediction equation : $y = b_0 + b_1x$



- We will start with the simple case of one **explanatory variable**.
- The **data vectors** are written as (x_i, y_i) for $i = 1, \ldots, n$.
- The simplest **prediction equation** to consider is a straight line that goes through the middle of the scatterplot of y versus x.

Chapter 2

- **Section 2.1** has the mathematics leading to the least squares line.

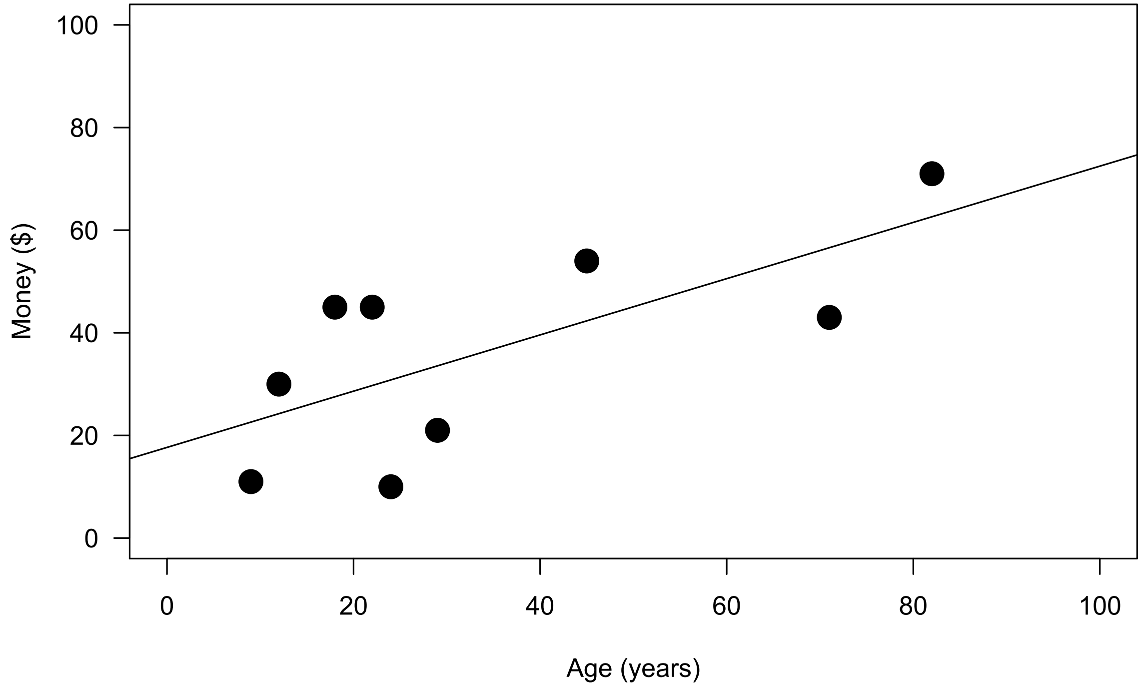
- **Section 2.2** introduces the simple linear regression model (prediction with one explanatory variable) that is formulated for a predictive equation. This is needed to quantify the variability of the coefficients of the best-fitting line, when different samples are taken from the population.

- **Section 2.5** has intervals for simple linear regression: the confidence interval for the slope of the least square line, confidence intervals for subpopulation means, and prediction intervals for a future or out-of-sample Y given x^*.

- **Section 2.6** has an explanation of Student t quantiles used in the interval estimates.

Section 2.1

“finding a best-fitting line to a scatterplot of a Y variable versus an X variable”



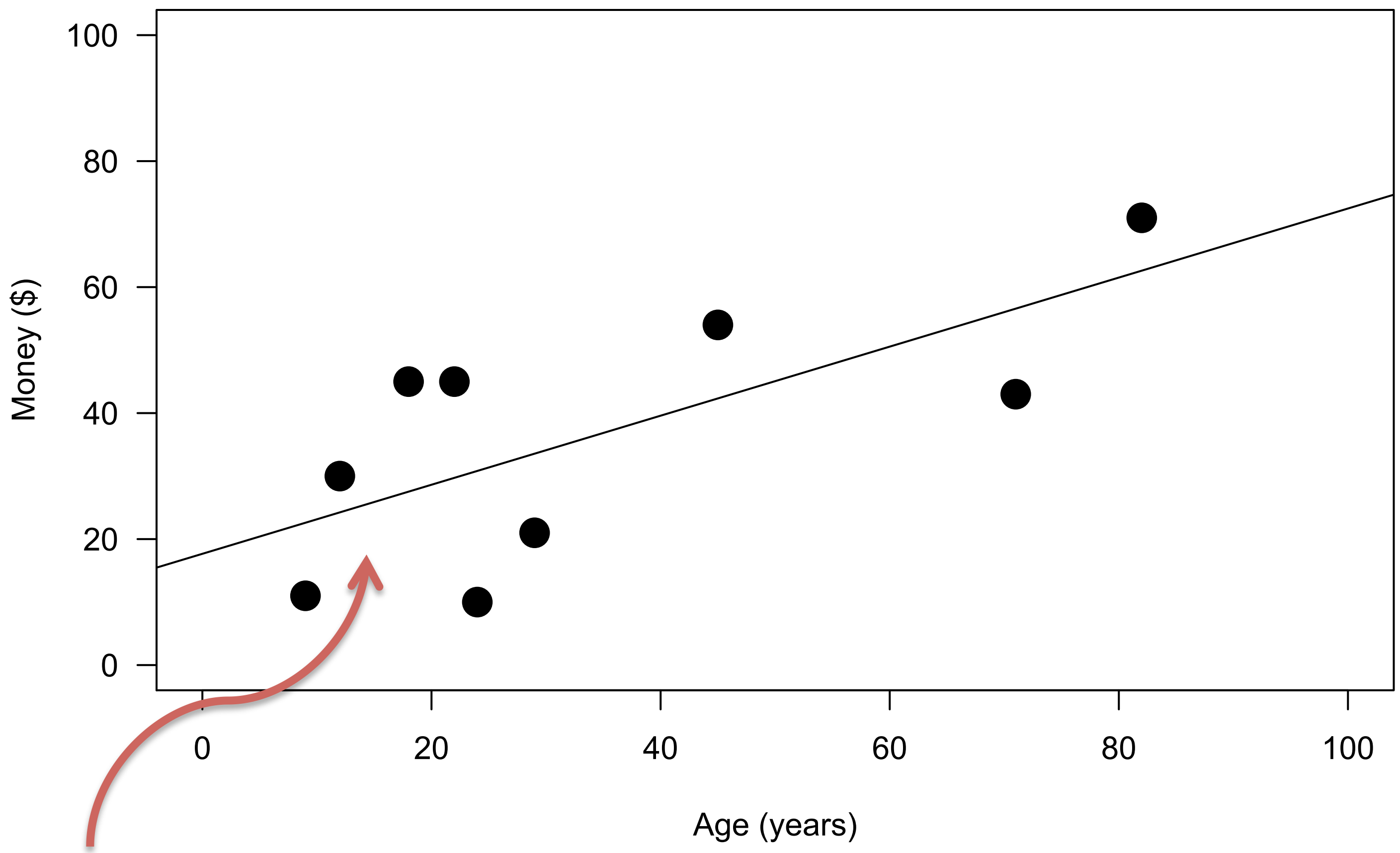
Section 2.1

"finding a best-fitting line to a scatterplot of a Y variable versus an X variable"

- 2.1.1 Visualization exercises
- 2.1.2 Summary statistics
- 2.1.3 Least squares solution

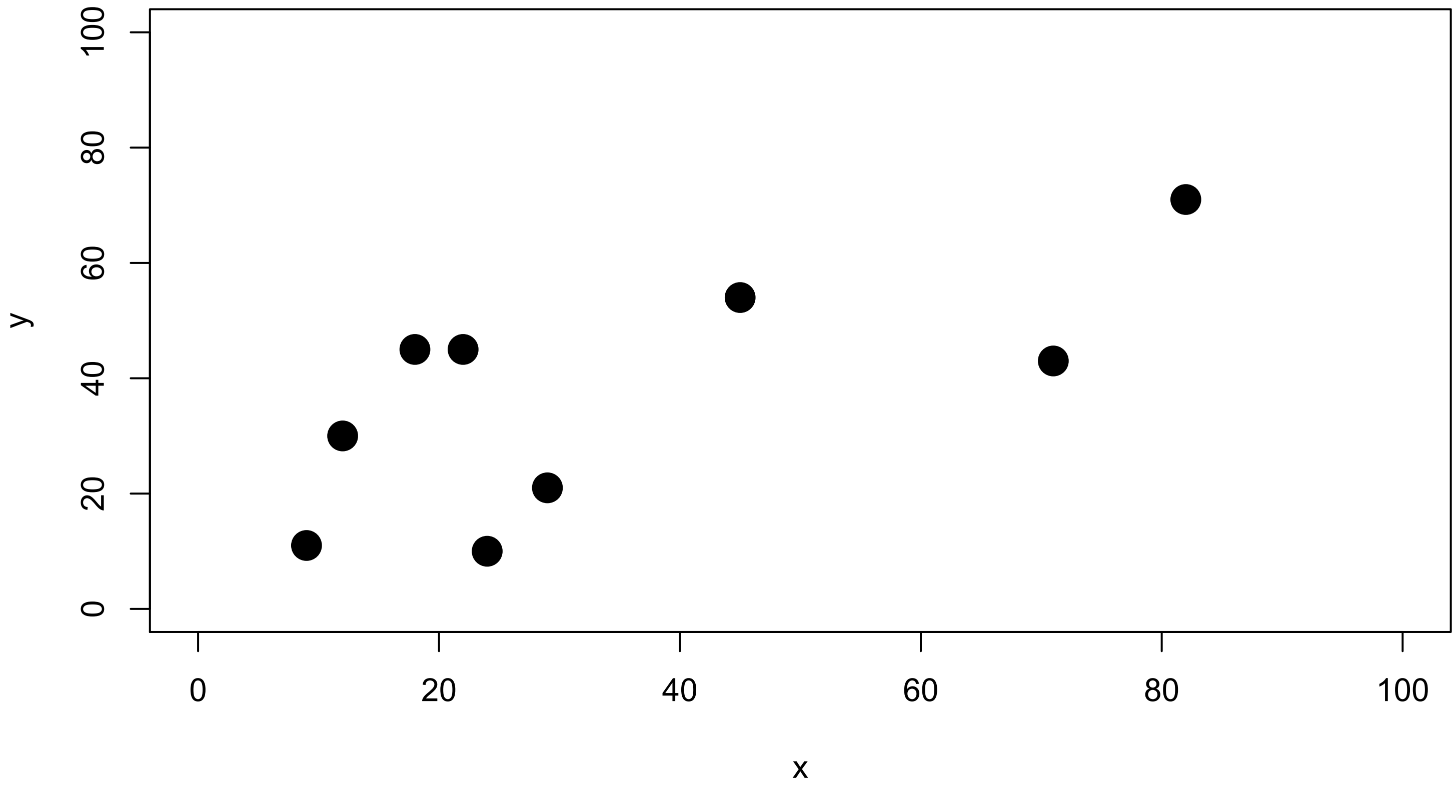
2.1.1 Visualization exercises

2.1.1 Visualization exercises



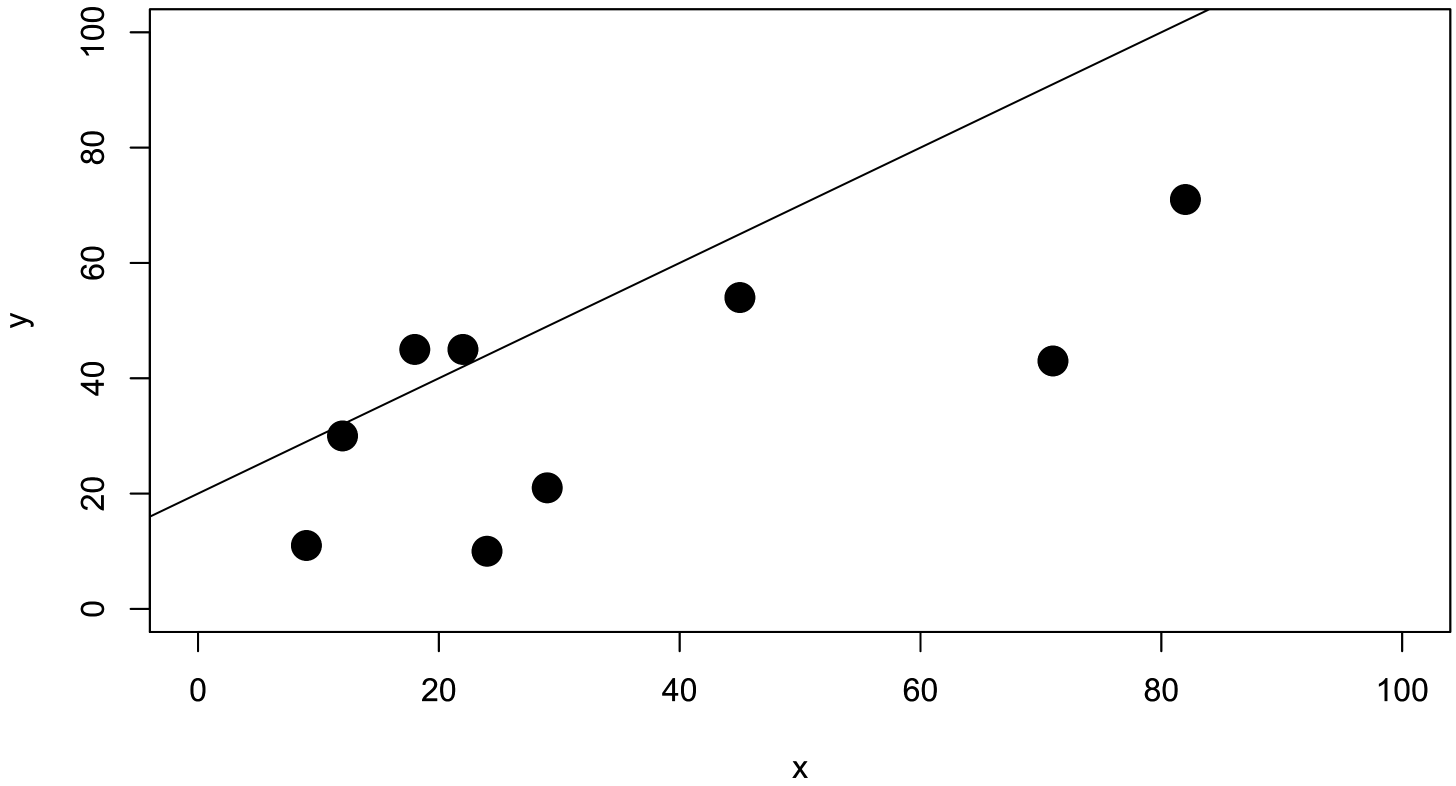
prediction equation

2.1.1 Visualization exercises



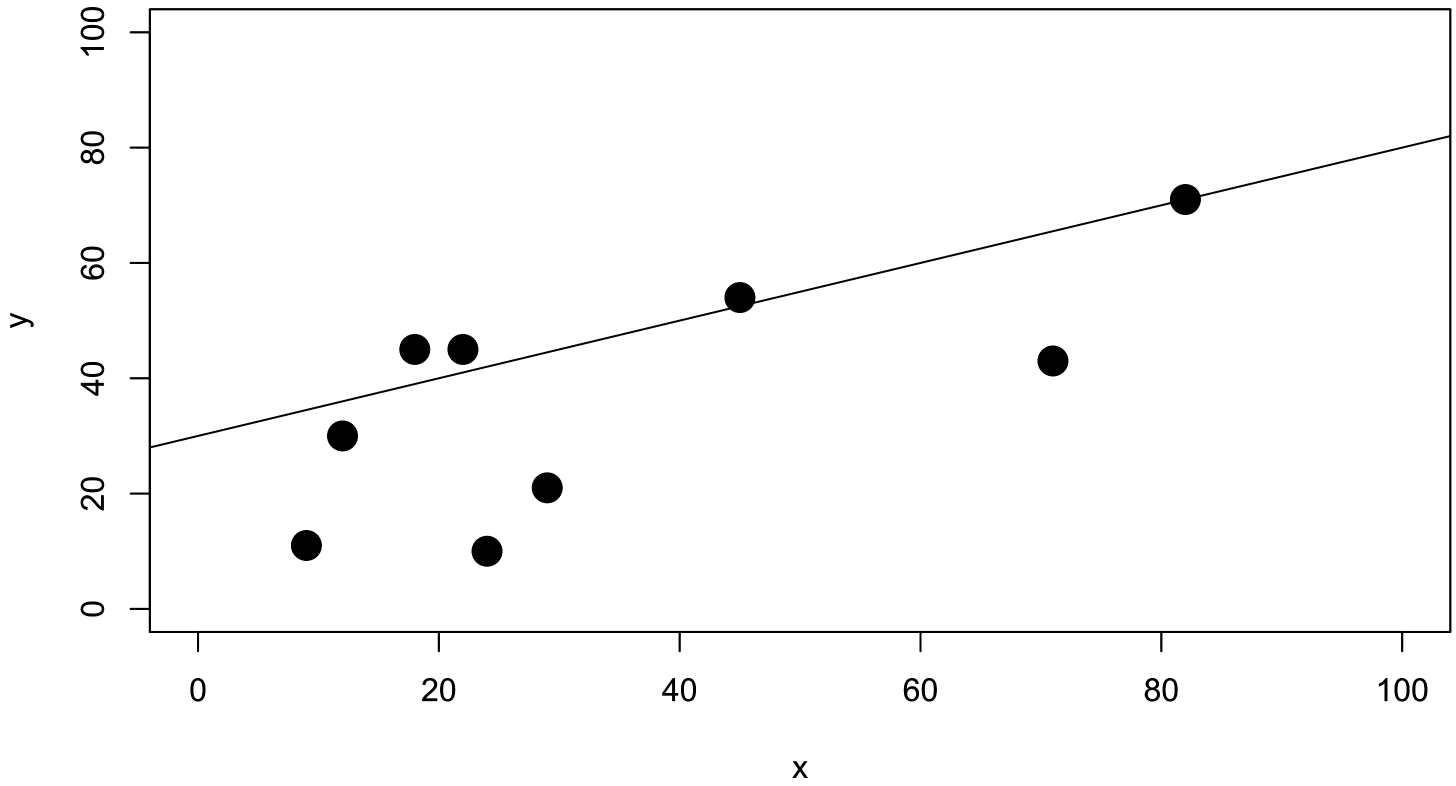

```
> x <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> n <- 9
> plot(y~x, pch=20, cex=3, xlim=c(0,100), ylim=c(0,100))
```

2.1.1 Visualization exercises



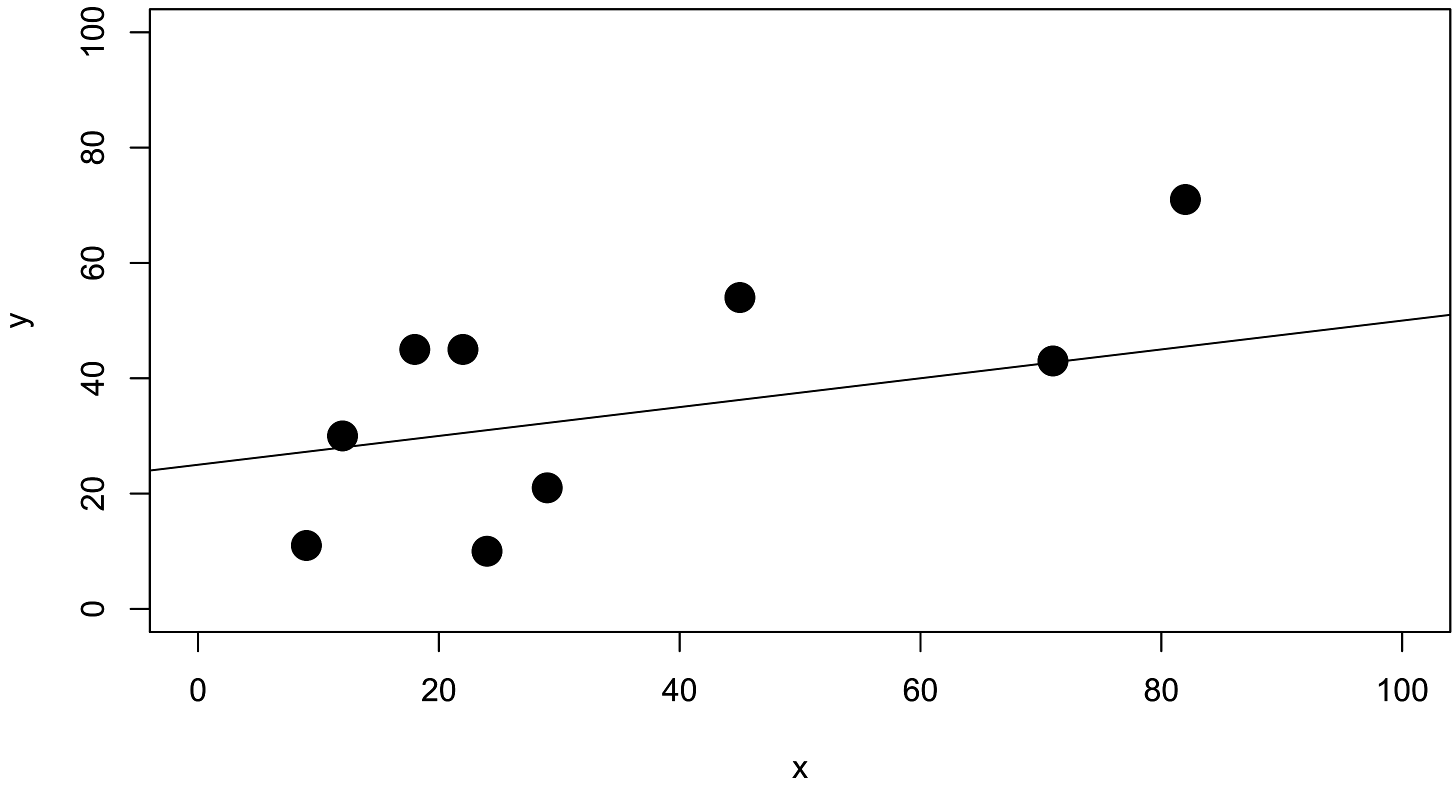

```
> x <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> n <- 9
> plot(y~x, pch=20, cex=3, xlim=c(0,100), ylim=c(0,100))
> abline(20,1)
```

2.1.1 Visualization exercises



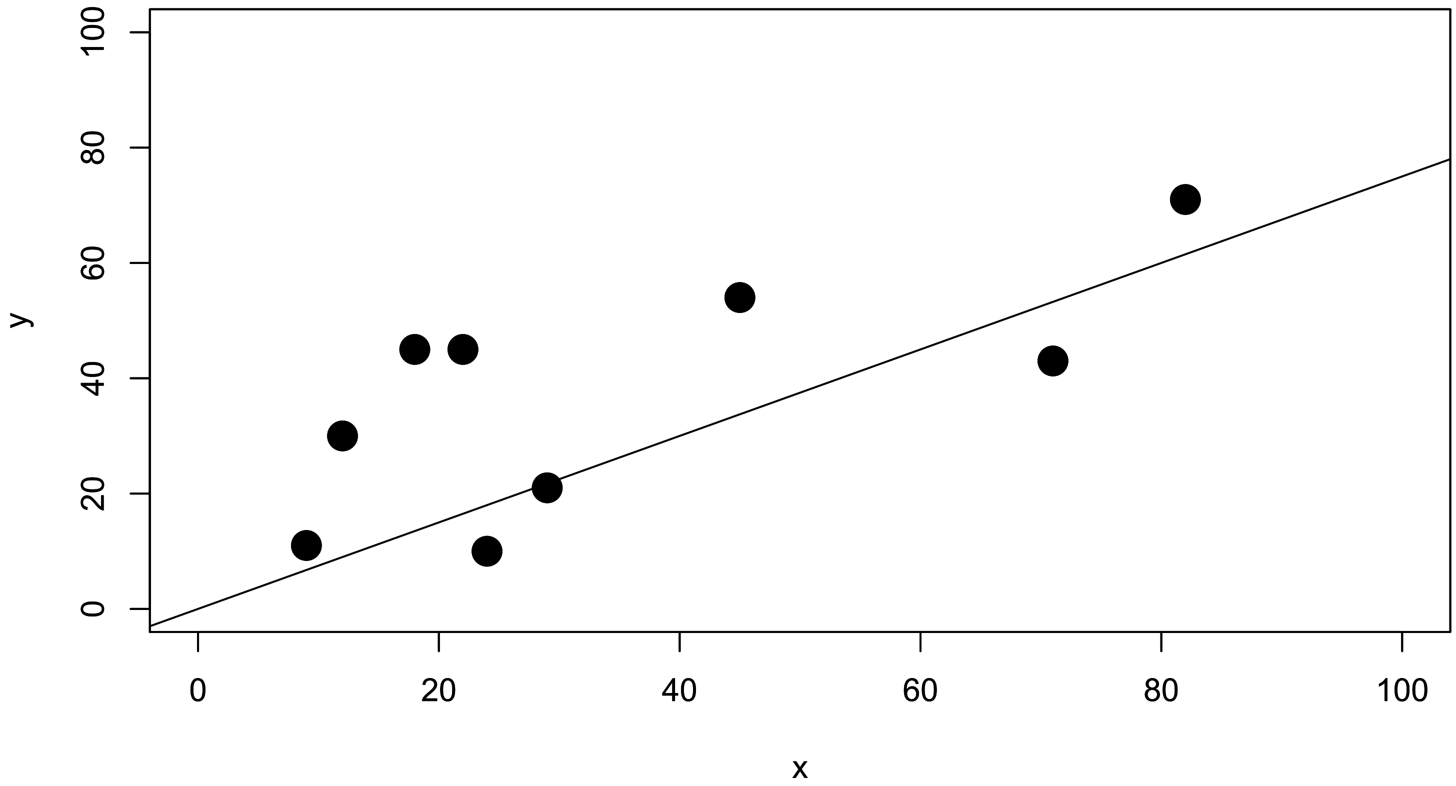

```
> x <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> n <- 9
> plot(y~x, pch=20, cex=3, xlim=c(0,100), ylim=c(0,100))
> abline(30,0.5)
```

2.1.1 Visualization exercises



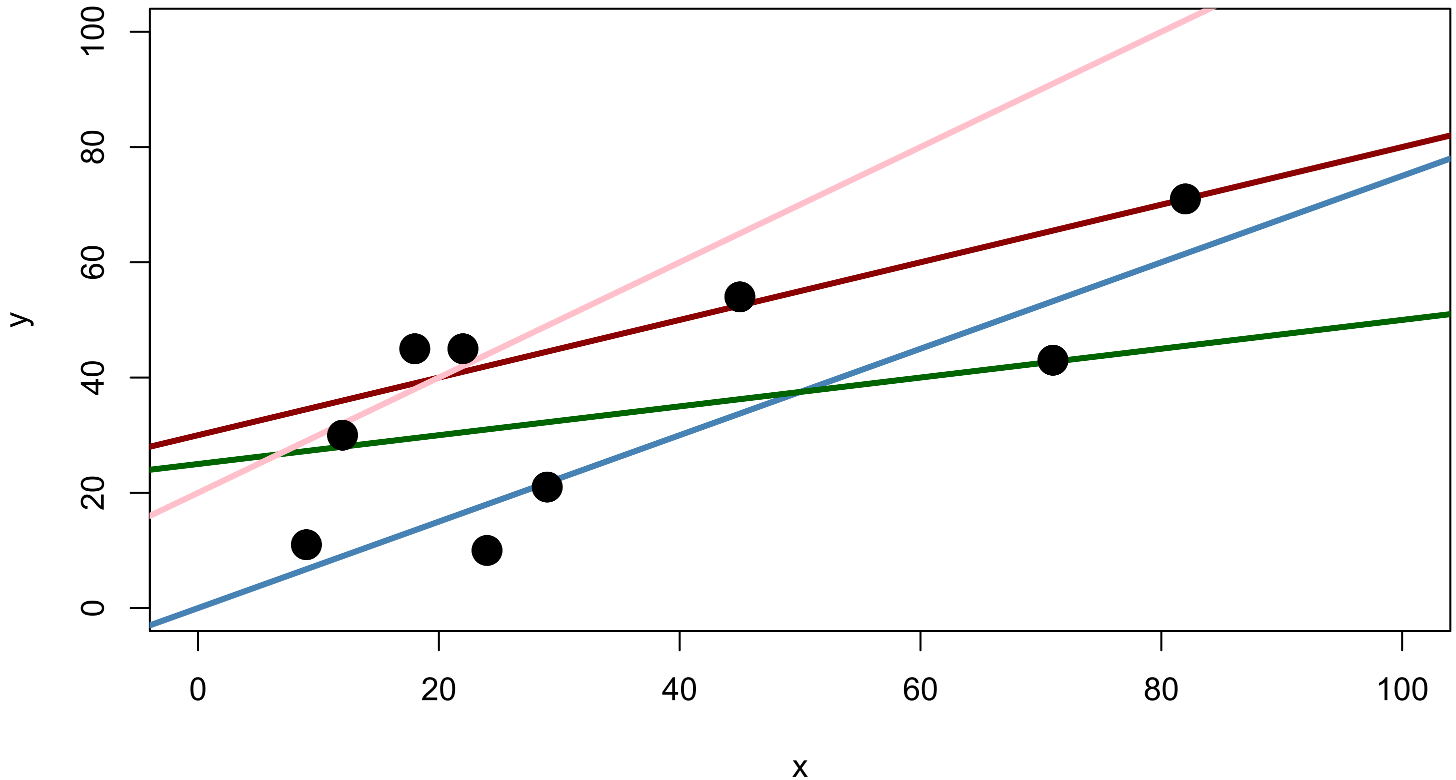

```
> x <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> n <- 9
> plot(y~x, pch=20, cex=3, xlim=c(0,100), ylim=c(0,100))
> abline(25,0.25)
```

2.1.1 Visualization exercises




```
> x <- c(82, 45, 71, 22, 29, 9, 12, 18, 24)
> y <- c(71, 54, 43, 45, 21, 11, 30, 45, 10)
> n <- 9
> plot(y~x, pch=20, cex=3, xlim=c(0,100), ylim=c(0,100))
> abline(0,0.75)
```

2.1.1 Visualization exercises



Which of the four lines (visually) looks like the best-fitting line?

$y = 0 + 0.75x$

$y = 25 + 0.25x$

$y = 30 + 0.5x$

$y = 20 + 1x$

Guess the regression line:

- https://www.geogebra.org/m/JsFmFEg6

- https://www.geogebra.org/m/B7JtA6Mg

2.1.2 Sample statistics

2.1.2 Sample statistics

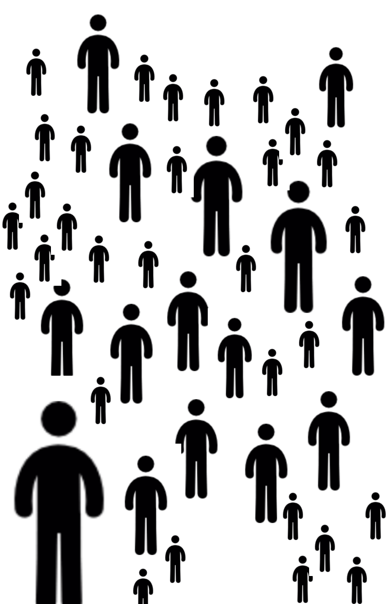
PREDICTOR variable

$X \longrightarrow$ Age in Years

RESPONSE variable

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$\beta_0, \beta_1, \sigma^2$

Hypothesis Test

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

Sample statistics

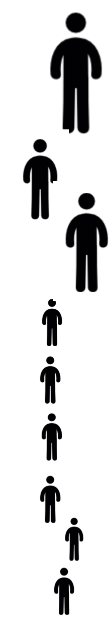
$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

For parameter β_1 :

95% C.I. = [0.05, 1.05]

p-value = 0.036

Sample, n=9



x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

2.1.2 Sample statistics

The data are first summarized into sample means (measures of centre)

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$$

$$\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$$

```
> xbar<-(1/n)*sum(x)
> xbar
[1] 34.66667
```

```
> ybar<-(1/n)*sum(y)
> ybar
[1] 36.66667
```

2.1.2 Sample statistics

The data are first summarized into sample means (measures of centre) and standard deviations (measures of spread):

$$\bar{x} = n^{-1}\sum_{i=1}^{n} x_i, \quad s_x = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \,/\, (n-1)}\,,$$

$$\bar{y} = n^{-1}\sum_{i=1}^{n} y_i, \quad s_y = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \,/\, (n-1)}\,.$$

```
> xbar<-(1/n)*sum(x)
> xbar
[1] 34.66667
>
> sx<-sqrt( sum((x-xbar)^2)/(n-1) )
> sx
[1] 26.03843
```

```
> ybar<-(1/n)*sum(y)
> ybar
[1] 36.66667
>
> sy<-sqrt( sum((y-ybar)^2)/(n-1) )
> sy
[1] 20.36541
```

2.1.2 Sample statistics

Correlation describes the linear association between X and Y.

- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases

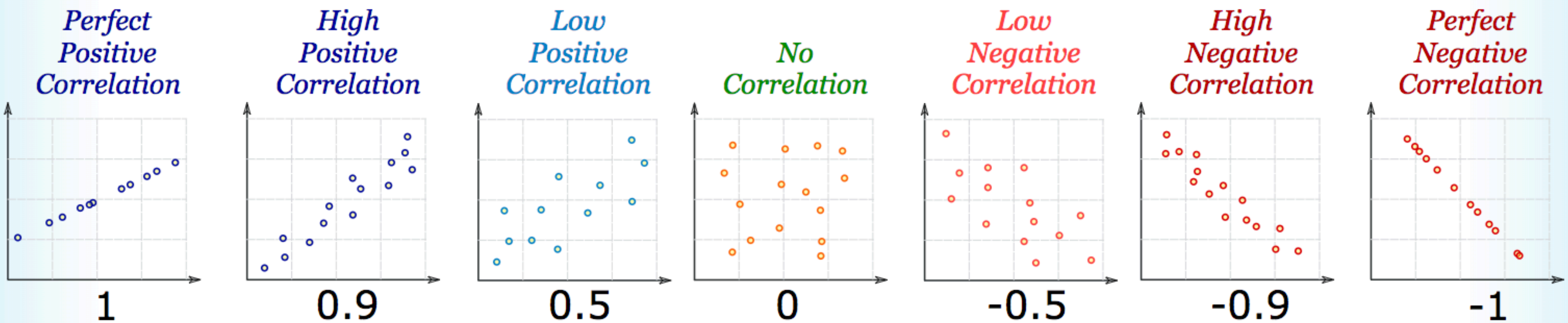
https://www.mathsisfun.com/data/correlation.html

Guess the Correlation Game: http://guessthecorrelation.com/

2.1.2 Sample statistics

- Correlation is **Positive** when the values **increase** together, and
- Correlation is **Negative** when one value **decreases** as the other increases

Here we look at **linear correlations** (correlations that follow a line).



Correlation can have a value:

- 1 is a perfect positive correlation
- 0 is no correlation (the values don't seem linked at all)
- -1 is a perfect negative correlation

https://www.mathsisfun.com/data/correlation.html

Guess the Correlation Game: **http://guessthecorrelation.com/**

2.1.2 Sample statistics

To summarize the linear association, the sample correlation is

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where sample covariance is

$$s_{xy} = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

2.1.2 Sample statistics

To summarize the linear association, the sample correlation is $r_{xy} = \frac{s_{xy}}{s_x s_y}$, where sample covariance is

$$s_{xy} = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \tag{2.3}$$

sample covariance:

```
> sxy<-(1/(n-1))*sum((x-xbar)*(y-ybar))
> sxy
[1] 371.625
```

2.1.2 Sample statistics

To summarize the linear association, the sample correlation is $r_{xy} = \frac{s_{xy}}{s_x s_y}$, where sample covariance is

$$s_{xy} = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}). \tag{2.3}$$

sample covariance:

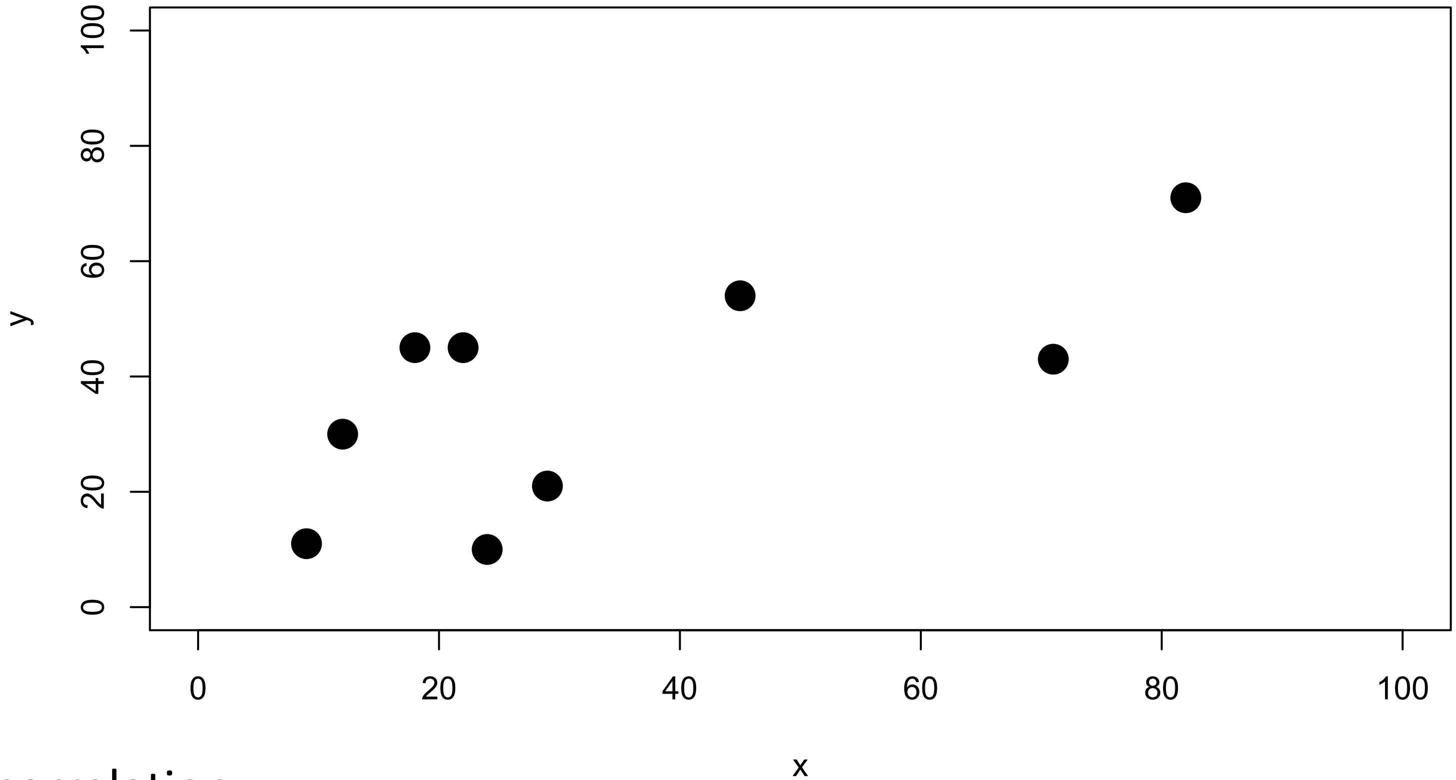
```
> sxy<-(1/(n-1))*sum((x-xbar)*(y-ybar))
> sxy
[1] 371.625
```

Sample correlation:

```
> rxy<-sxy/(sx*sy)
> rxy
[1] 0.7008045
```

Note that $-1 \leq r_{xy} \leq 1$ is a scaled version of the unbounded s_{xy}, that is, $-\infty < s_{xy} < \infty$. The quantities r_{xy} and s_{xy} are positive (negative) if the scatterplot of (x_i, y_i), $i = 1, \ldots, n$, form a cloud the slopes upwards (downwards).

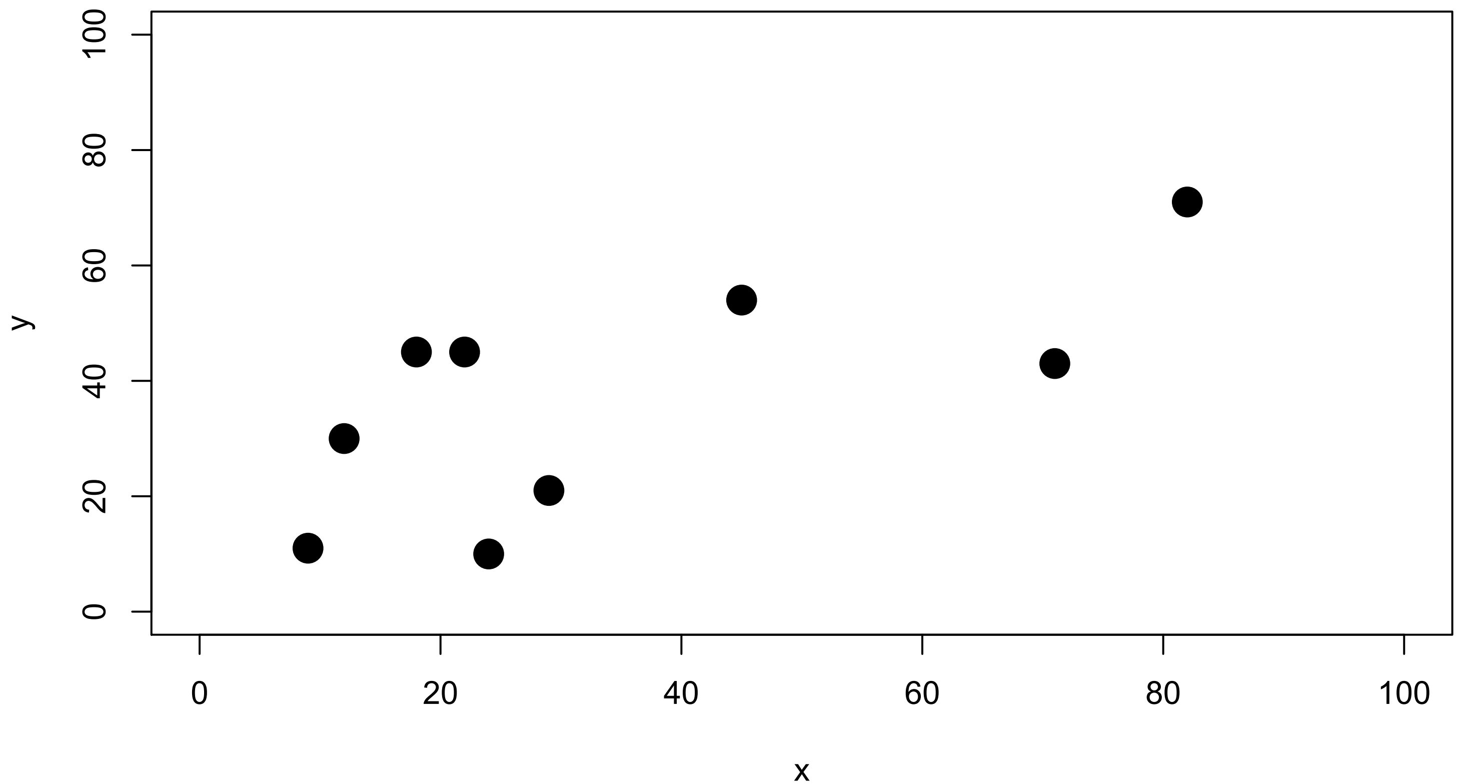
2.1.2 Sample statistics



Sample correlation:

```
> rxy<-sxy/(sx*sy)
> rxy
[1] 0.7008045
```

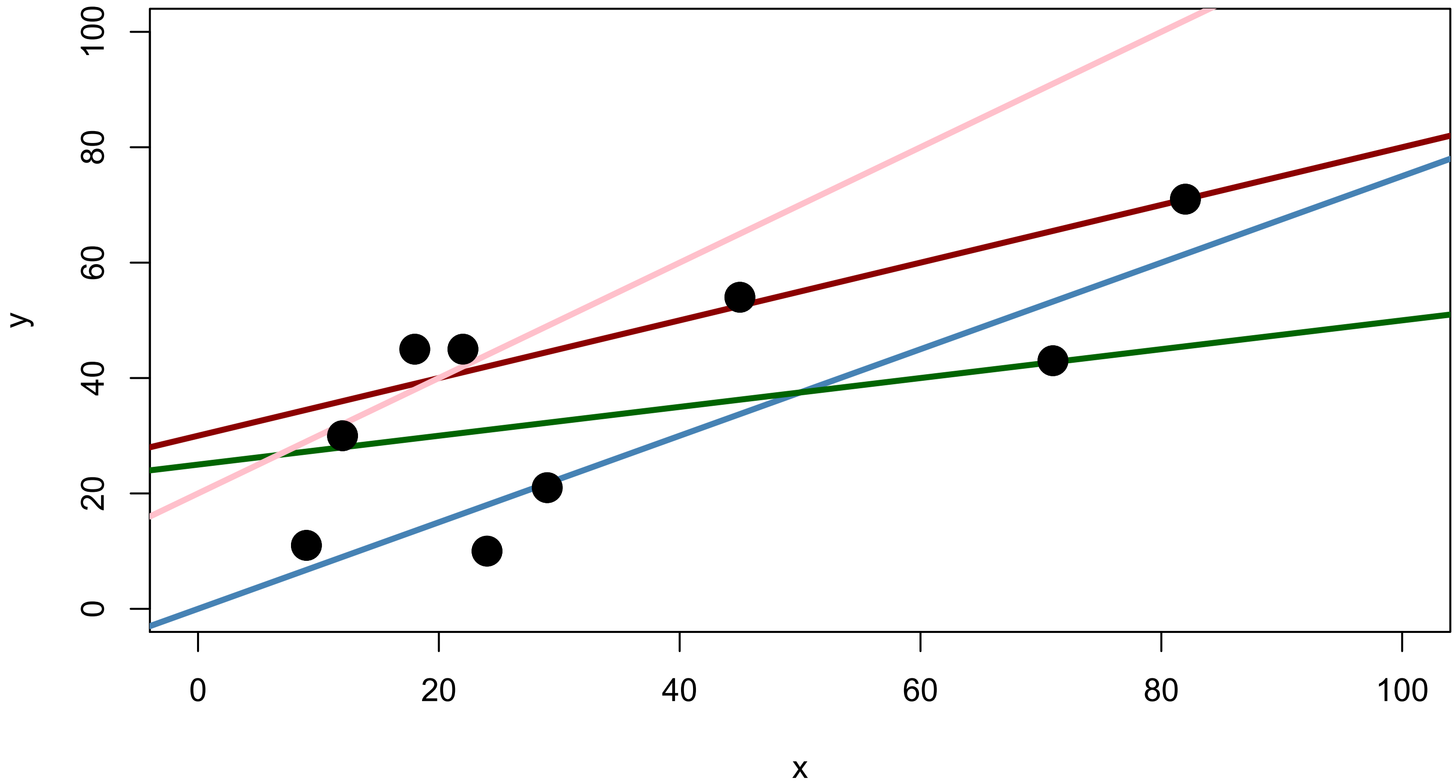
2.1.2 Sample statistics



Data don't lie on a straight line or smooth curve because there is noise or variation from other factors that affect the response.

If the two variables are linearly related, a summary is a line $y = b_0 + b_1x$ that goes through the middle of the scatterplot.

2.1.1 Visualization exercises



Which of the four lines (visually) looks like the best-fitting line?

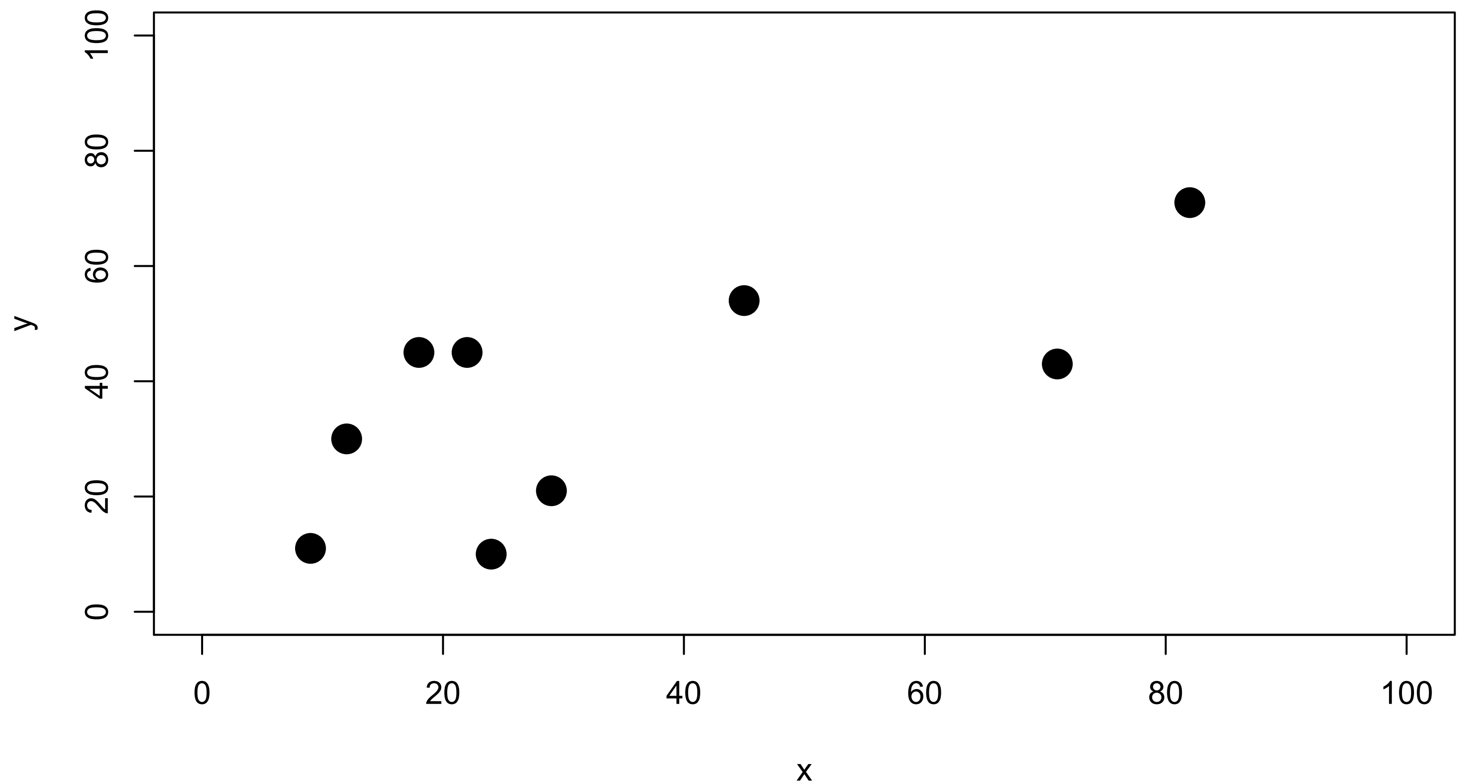
$y = 0 + 0.75x$

$y = 25 + 0.25x$

$y = 30 + 0.5x$

$y = 20 + 1x$

2.1.2 Sample statistics



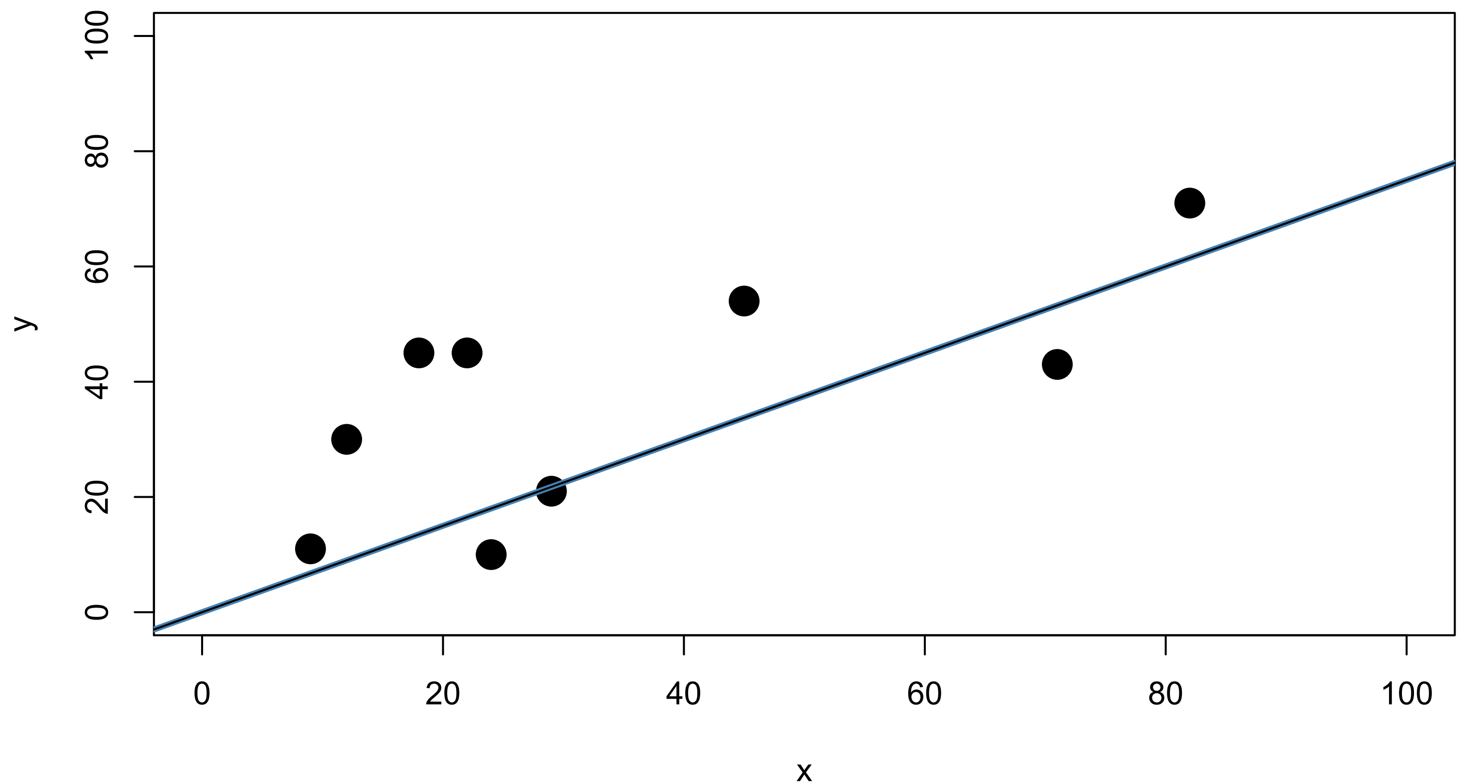
To find b_0 and b_1 of a best-fitting line, one choice is to minimize **the Sum of Squared Errors**:

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

What is the Sum of Squared Errors? https://www.geogebra.org/m/JsFmFEg6

or http://students.brown.edu/seeing-theory/regression/index.html

2.1.2 Sample statistics

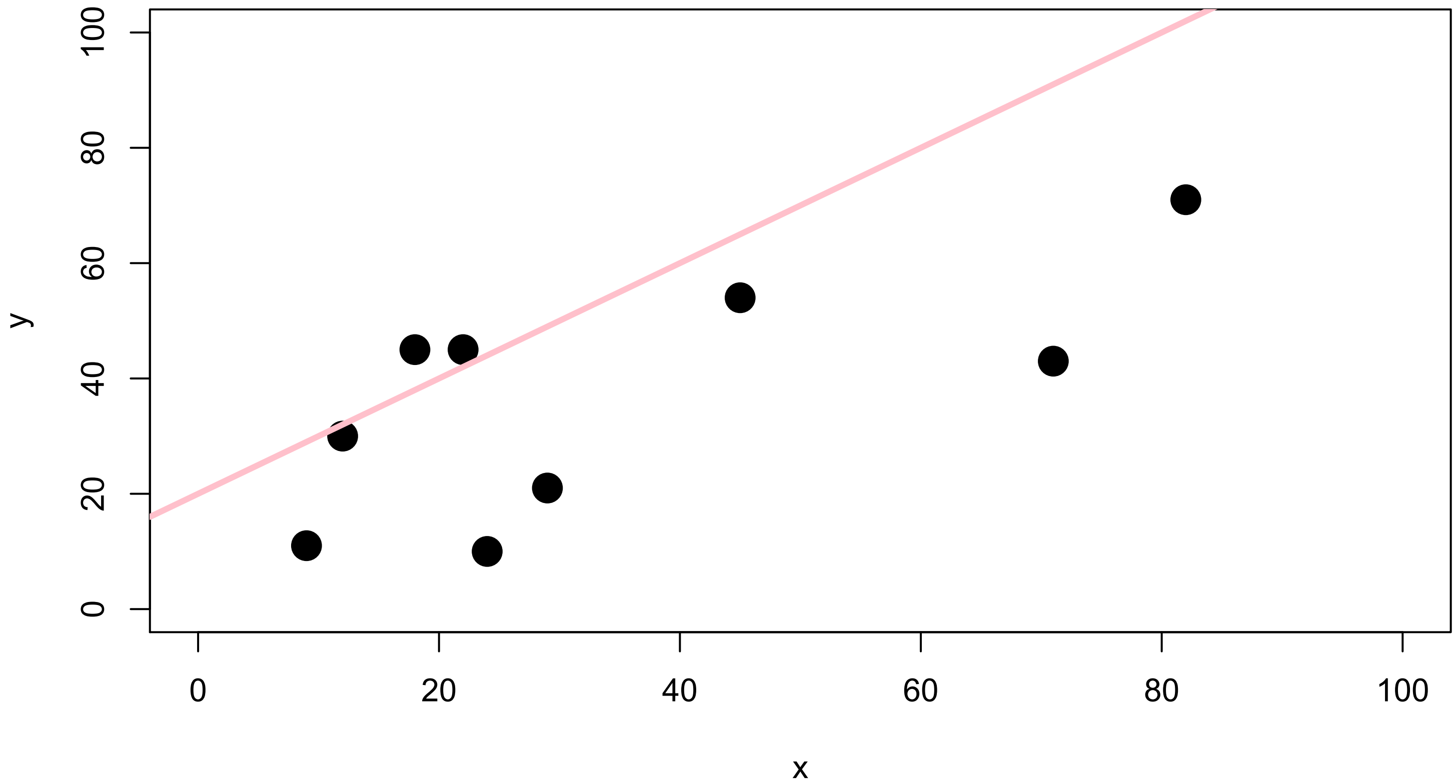


To find b_0 and b_1 of a best-fitting line, one choice is to minimize the **SSE**:

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

```
> b0<-0; b1<-0.75
> sum( (y- b0 - b1*x)^2)
[1] 2933.5
```

2.1.2 Sample statistics

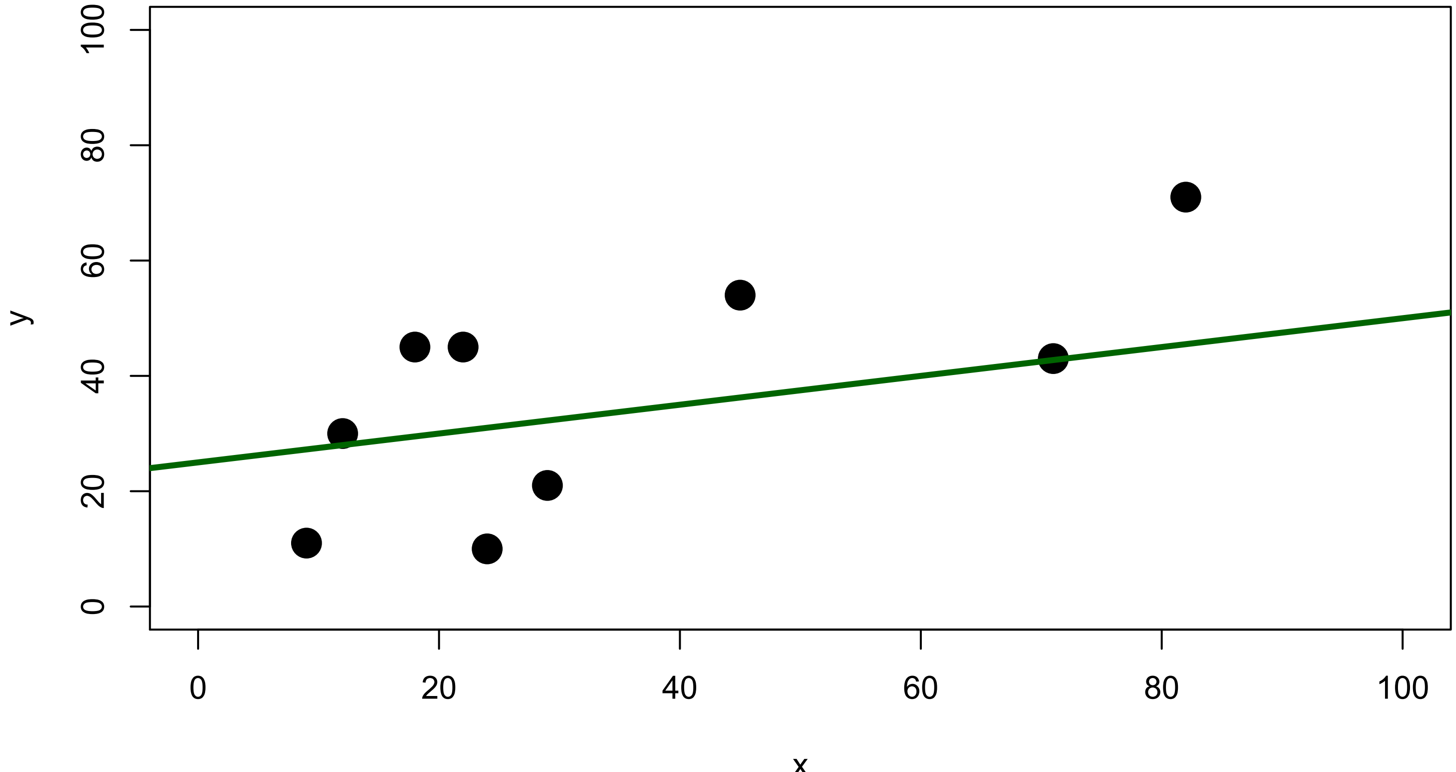


To find b_0 and b_1 of a best-fitting line, one choice is to minimize:

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

```
> b0<-20; b1<-1
> sum( (y- b0 - b1*x)^2)
[1] 5712
```

2.1.2 Sample statistics

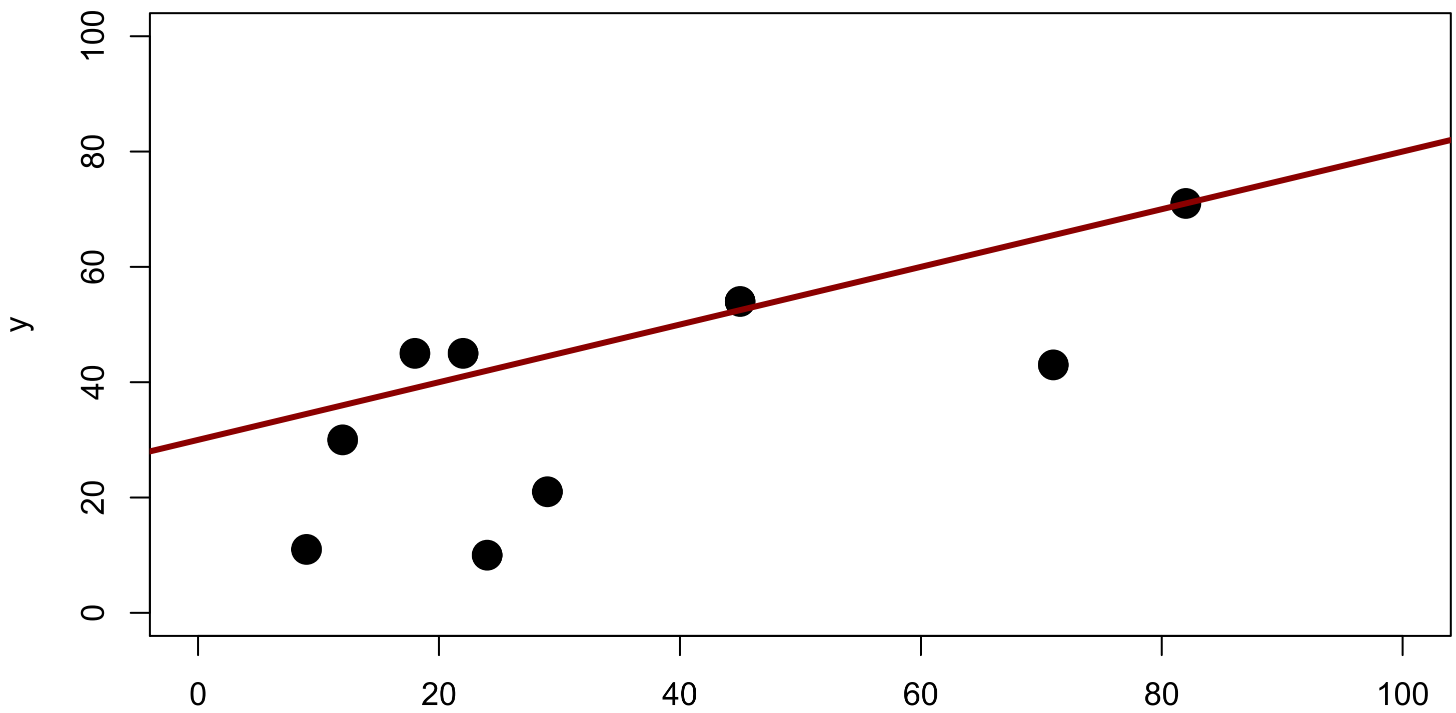


To find b_0 and b_1 of a best-fitting line, one choice is to minimize:

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

```
> b0<-25; b1<-0.25
> sum( (y- b0 - b1*x)^2)
[1] 2251.5
```

2.1.2 Sample statistics

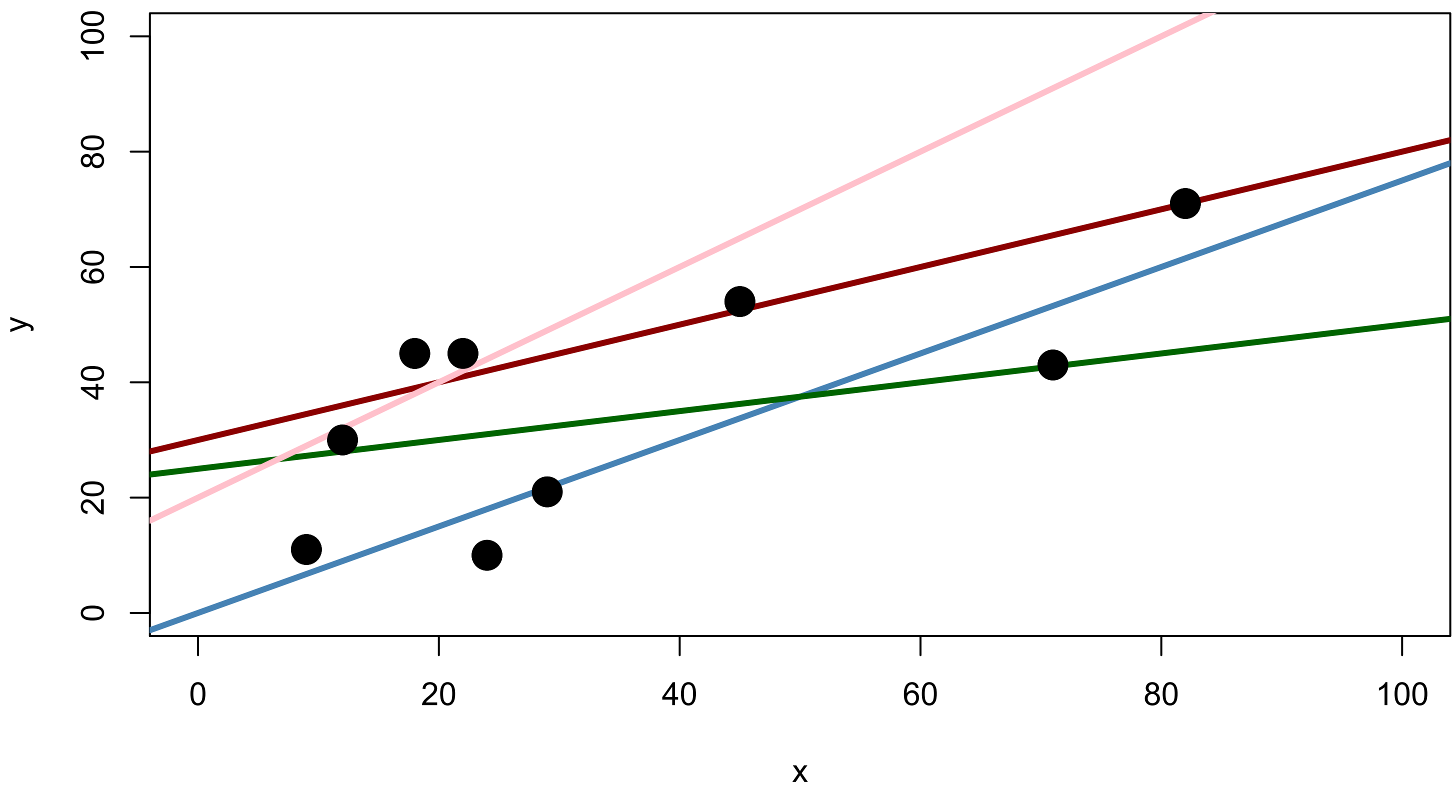


To find b_0 and b_1 of a best-fitting line, one choice is to minimize:

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

```
> b0<-30; b1<-0.5
> sum( (y- b0 - b1*x)^2)
[1] 2725
```

2.1.2 Sample statistics



Which of the four lines (visually) looks like the best-fitting line?

$y = 0 + 0.75x$	$S(b0,b1) = 2933.5$
$y = 25 + 0.25x$	$S(b0,b1) = 2251.5$
$y = 30 + 0.5x$	$S(b0,b1) = 2725.0$
$y = 20 + 1x$	$S(b0,b1) = 5712.0$

2.1.2 Sample statistics

To find b_0 and b_1 of a best-fitting line, one choice is to minimize:

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

The line that minimizes $S(b_0,b_1)$ is called the:

least squares regression line (LSRL)

2.1.2 Sample statistics

To find b_0 and b_1 of a best-fitting line, one choice is to minimize:

$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

The line that minimizes $S(b_0, b_1)$ is called the:

least squares regression line (LSRL)

It minimizes the **sum of squared errors (also known as: "SSE" or "squared vertical deviations" or "Sum of Squares")**.

Question: How do we find the values of b_0 and b_1 that minimize $S(b_0, b_1)$?

Answer: Simple calculus

2.1.3 Least squares solution

2.1.3 Least squares solution

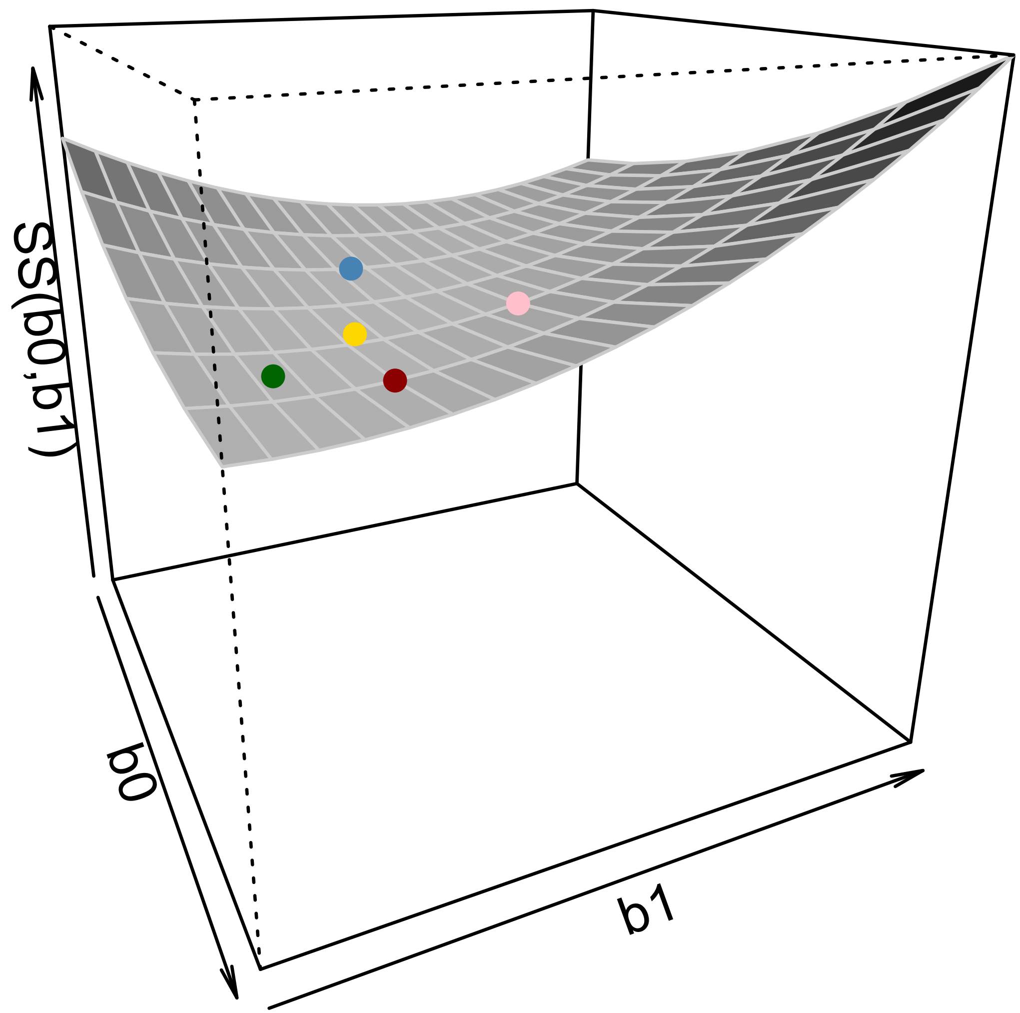
$$S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

The goal is to minimize $S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$.

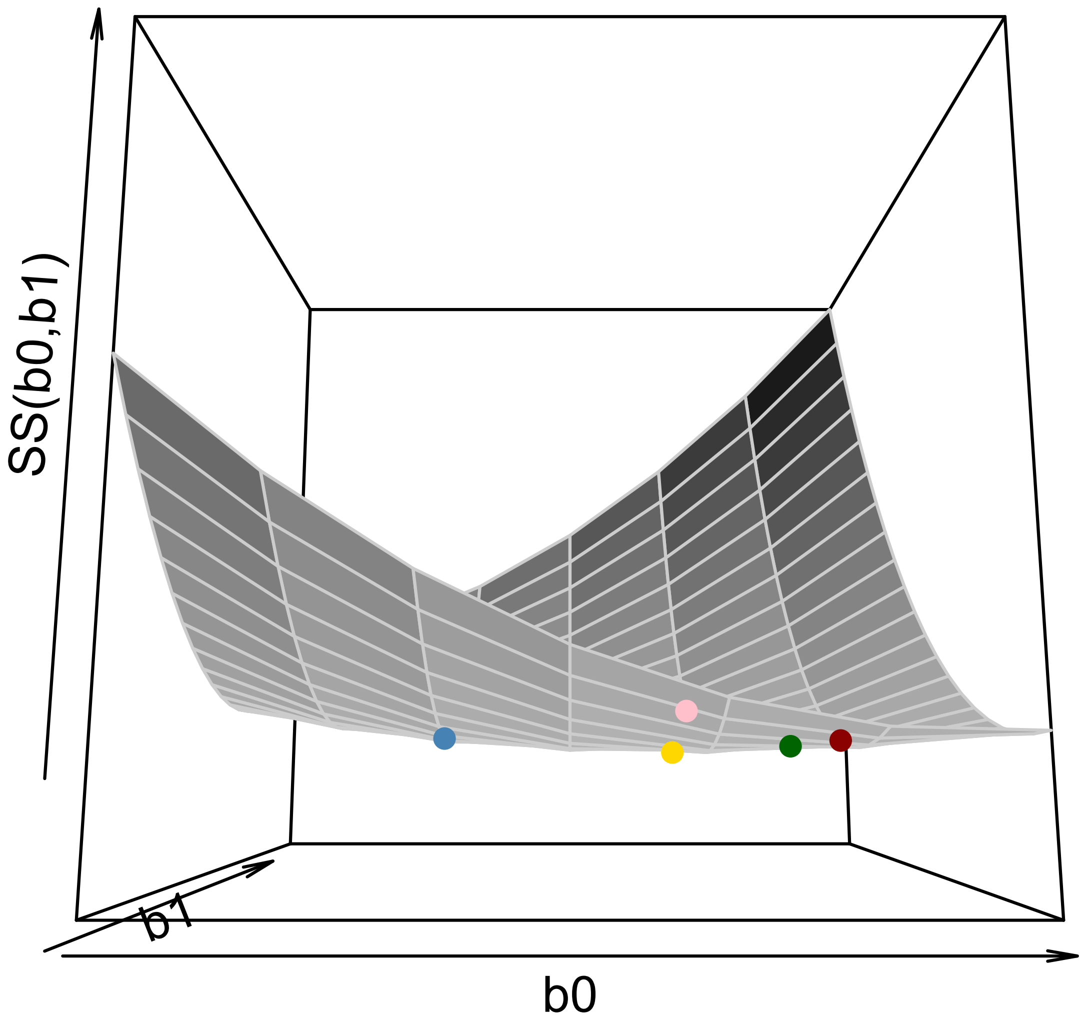
The partial derivatives of S are:

$$\begin{aligned} \frac{\partial S}{\partial b_0} &= -2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i), \\ \frac{\partial S}{\partial b_1} &= -2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i). \end{aligned}$$

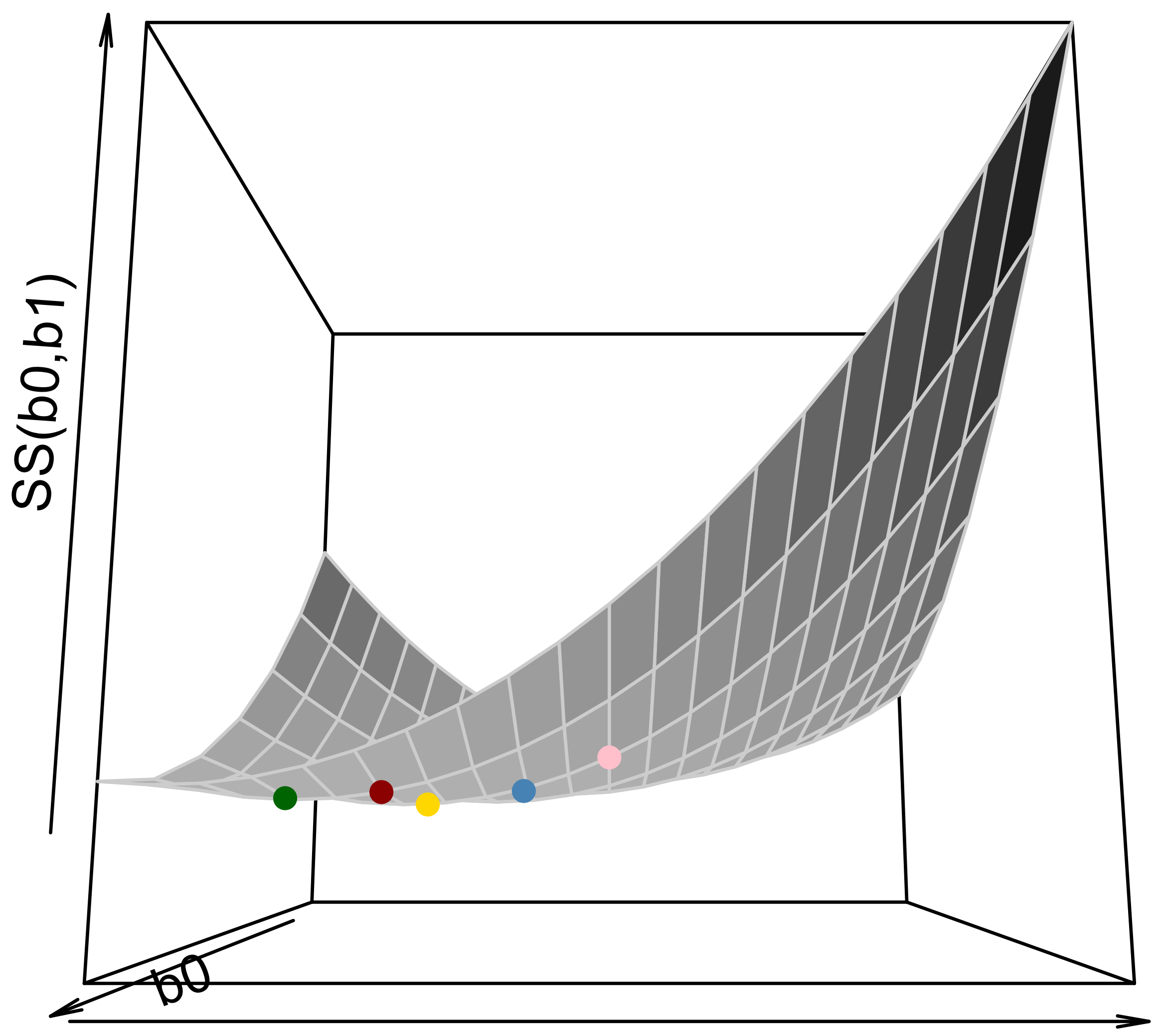
2.1.3 Least squares solution



2.1.3 Least squares solution



2.1.3 Least squares solution



2.1.3 Least squares solution

The goal is to minimize $S(b_0, b_1) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$.

$$
\begin{aligned}
\frac{\partial S}{\partial b_0} &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i), \\
\frac{\partial S}{\partial b_1} &= -2\sum_{i=1}^{n}x_i(y_i - b_0 - b_1 x_i).
\end{aligned}
$$

Set the equations to 0, divide by -2 and solve.

TIP: Go through this process with a pen and paper until you can easily obtain the solution yourself.

2.1.3 Least squares solution

The goal is to minimize $S(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$.

Set the equations to 0, divide by -2 and solve.

The solution $(\hat{b}_0, \hat{b}_1)$ satisfies

$$
\begin{aligned}
0 &= n[\overline{y} - \hat{b}_0 - \hat{b}_1 \overline{x}], \\
0 &= \sum_{i=1}^{n} x_i y_i - \hat{b}_0 n \overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2.
\end{aligned}
$$

2.1.3 Least squares solution

$$
\begin{aligned}
0 &= n[\overline{y} - \hat{b}_0 - \hat{b}_1\overline{x}], \\
0 &= \sum_{i=1}^{n} x_i y_i - \hat{b}_0 n\overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2.
\end{aligned}
$$

Further simplify with

$$
\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}, \tag{2.18}
$$

2.1.3 Least squares solution

$$
\begin{aligned}
0 &= n[\overline{y} - \hat{b}_0 - \hat{b}_1 \overline{x}], \\
0 &= \sum_{i=1}^{n} x_i y_i - \hat{b}_0 n \overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2.
\end{aligned}
$$

Further simplify with

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}, \tag{2.18}$$

$$0 = \sum_{i=1}^{n} x_i y_i - [\overline{y} - \hat{b}_1 \overline{x}] n \overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.19}$$

2.1.3 Least squares solution

$$
\begin{aligned}
0 &= n[\overline{y} - \hat{b}_0 - \hat{b}_1 \overline{x}], \\
0 &= \sum_{i=1}^{n} x_i y_i - \hat{b}_0 n \overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2.
\end{aligned}
$$

Further simplify with

$$
\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}, \tag{2.18}
$$

$$
0 = \sum_{i=1}^{n} x_i y_i - [\overline{y} - \hat{b}_1 \overline{x}] n \overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.19}
$$

$$
0 = \sum_{i=1}^{n} x_i y_i - n \overline{x}\overline{y} + n \hat{b}_1 \overline{x}^2 - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.20}
$$

2.1.3 Least squares solution

Further simplify with

$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}, \tag{2.18}$$

$$0 = \sum_{i=1}^{n} x_i y_i - [\overline{y} - \hat{b}_1\overline{x}]n\overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.19}$$

$$0 = \sum_{i=1}^{n} x_i y_i - n\overline{xy} + n\hat{b}_1\overline{x}^2 - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.20}$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{2.21}$$

2.1.3 Least squares solution

Further simplify with

$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}, \tag{2.18}$$

$$0 = \sum_{i=1}^{n} x_i y_i - [\overline{y} - \hat{b}_1\overline{x}]n\overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.19}$$

$$0 = \sum_{i=1}^{n} x_i y_i - n\overline{xy} + n\hat{b}_1\overline{x}^2 - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.20}$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{2.21}$$

$$= \frac{(n-1)s_{xy}}{(n-1)s_x^2} \tag{2.22}$$

2.1.3 Least squares solution

Further simplify with

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}, \tag{2.18}$$

$$0 = \sum_{i=1}^{n} x_i y_i - [\overline{y} - \hat{b}_1 \overline{x}] n \overline{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.19}$$

$$0 = \sum_{i=1}^{n} x_i y_i - n \overline{x} \overline{y} + n \hat{b}_1 \overline{x}^2 - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.20}$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n \overline{x} \overline{y}}{\sum_{i=1}^{n} x_i^2 - n \overline{x}^2} \tag{2.21}$$

$$= \frac{(n-1) s_{xy}}{(n-1) s_x^2} \tag{2.22}$$

$$= \frac{r_{xy} s_x s_y}{s_x^2} = \frac{r_{xy} s_y}{s_x}. \tag{2.23}$$

2.1.3 Least squares solution

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}, \tag{2.18}$$

$$0 = \sum_{i=1}^{n} x_i y_i - [\bar{y} - \hat{b}_1\bar{x}]n\bar{x} - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.19}$$

$$0 = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} + n\hat{b}_1\bar{x}^2 - \hat{b}_1 \sum_{i=1}^{n} x_i^2, \tag{2.20}$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{2.21}$$

$$= \frac{(n-1)s_{xy}}{(n-1)s_x^2} \tag{2.22}$$

$$= \frac{r_{xy}s_x s_y}{s_x^2} = \frac{r_{xy}s_y}{s_x}. \tag{2.23}$$

The solution is therefore:

$$\hat{b}_0 = \bar{y} - \hat{b}_1\bar{x}$$

$$\hat{b}_1 = r_{xy}s_y/s_x$$

2.1.3 Least squares solution

The solution is therefore:

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

$$\hat{b}_1 = r_{xy} s_y / s_x$$

```
> xbar<-(1/n)*sum(x)
> xbar
[1] 34.66667
>
> sx<-sqrt( sum((x-xbar)^2)/(n-1) )
> sx
[1] 26.03843
>
> ybar<-(1/n)*sum(y)
> ybar
[1] 36.66667
>
> sy<-sqrt( sum((y-ybar)^2)/(n-1) )
> sy
[1] 20.36541
>
> sxy<-(1/(n-1))*sum((x-xbar)*(y-ybar))
> sxy
[1] 371.625
```

```
> rxy<-sxy/(sx*sy)
> rxy
[1] 0.7008045
>
> b1_hat<-rxy*sy/sx
> b0_hat<-ybar-b1_hat*xbar
>
> b1_hat
[1] 0.5481195
> b0_hat
[1] 17.66519
```

2.1.3 Least squares solution

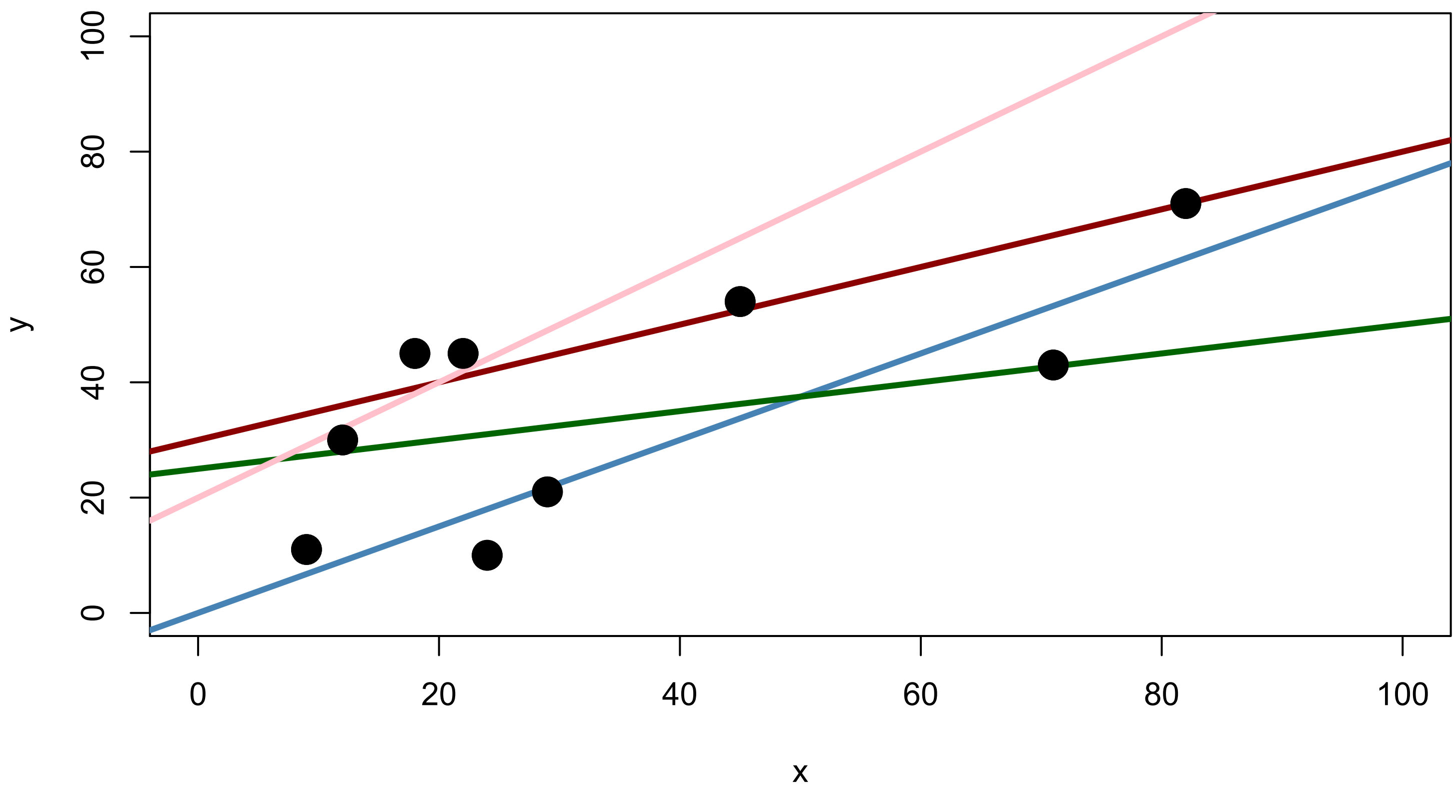
The solution is therefore:

$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}$$

$$\hat{b}_1 = r_{xy}s_y/s_x$$

```
> b1_hat<-rxy*sy/sx
> b0_hat<-ybar-b1_hat*xbar
>
> b1_hat
[1] 0.5481195
> b0_hat
[1] 17.66519
>
> b0<-17.665; b1<-0.548
> sum( (y- b0 - b1*x)^2)
[1] 1688.441
```

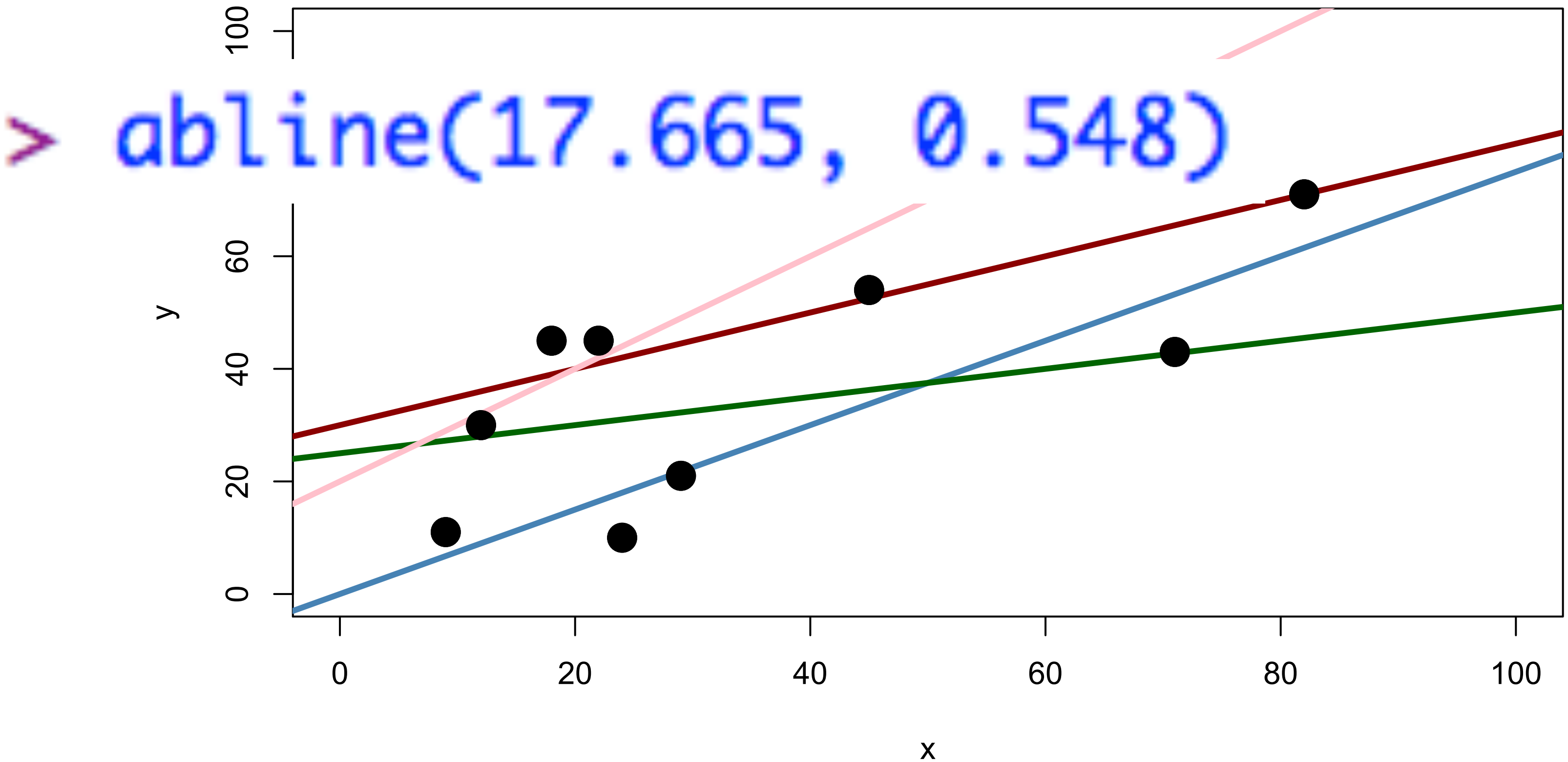
2.1.3 Least squares solution



Which of the four lines (visually) looks like the best-fitting line?

$y = 0 + 1x$	$S(b0,b1) = 2933.5$
$y = 25 + 0.25x$	$S(b0,b1) = 2251.5$
$y = 30 + 0.5x$	$S(b0,b1) = 2725.0$
$y = 20 + 1x$	$S(b0,b1) = 5712.0$

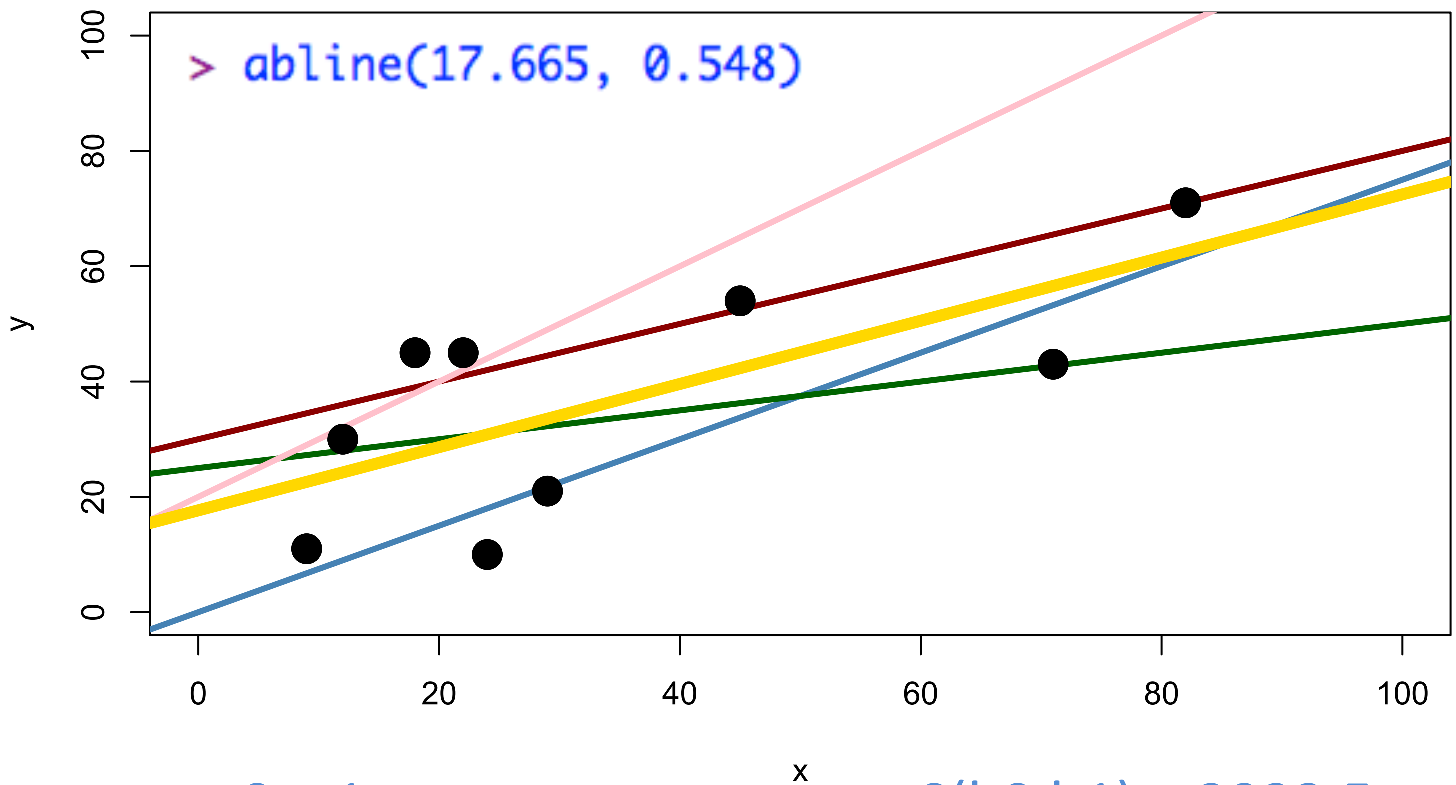
2.1.3 Least squares solution



Which of the four lines (visually) looks like the best-fitting line?

y = 0 + 1x	S(b0,b1) = 2933.5
y = 25 + 0.25x	S(b0,b1) = 2251.5
y = 30 + 0.5x	S(b0,b1) = 2725.0
y = 20 + 1x	S(b0,b1) = 5712.0

2.1.3 Least squares solution



y = 0 + 1x S(b0,b1) = 2933.5

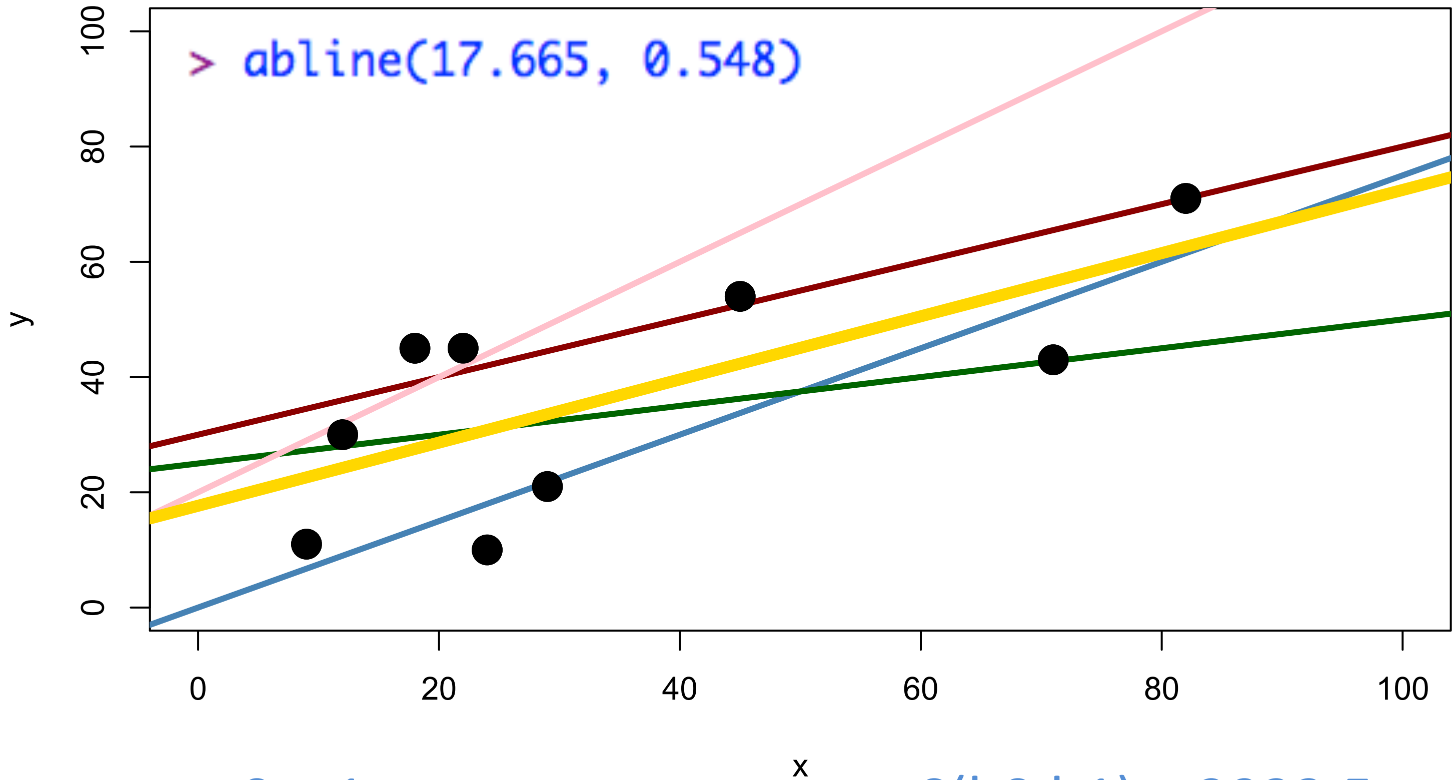
y = 25 + 0.25x S(b0,b1) = 2251.5

y = 30 + 0.5x S(b0,b1) = 2725.0

y = 20 + 1x S(b0,b1) = 5712.0

y = 17.665 + 0.548x S(b0,b1) = 1688.4

2.1.3 Least squares solution



$y = 0 + 1x$	$S(b0,b1) = 2933.5$
$y = 25 + 0.25x$	$S(b0,b1) = 2251.5$
$y = 30 + 0.5x$	$S(b0,b1) = 2725.0$
$y = 20 + 1x$	$S(b0,b1) = 5712.0$
$y = 17.7 + 0.55x$	$S(b0,b1) = 1688.4$

Age vs. Money

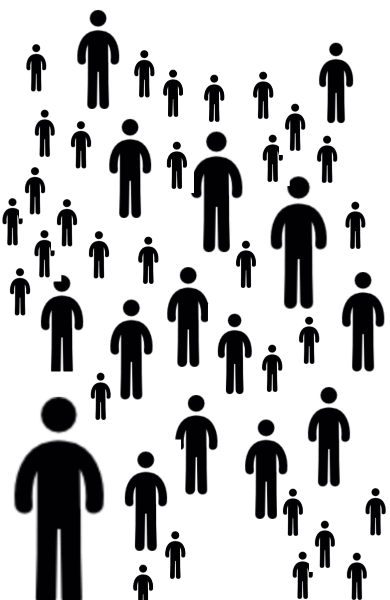
PREDICTOR variable

$X \longrightarrow$ Age in Years

RESPONSE variable

$Y \longrightarrow$ dollars ($) In bank account

Population



Population parameters

$$\beta_0, \beta_1, \sigma^2$$

Hypothesis Test

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

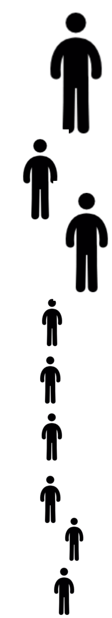
Sample statistics

b_0 = **17.7**
b_1 = **0.55**
s = 15.5
R^2 = 0.49

For parameter β_1 :

95% C.I. = [0.05, 1.05]
p-value = 0.036

Sample, n=9



x	y
82	71
45	54
71	43
22	45
29	21
9	11
12	30
18	45
24	10

Age vs. Money

Sample statistics

$b_0 = 17.7$
$b_1 = 0.55$
$s = 15.5$
$R^2 = 0.49$

Objective: The purpose of this observational study was to demonstrate if, and to what extent, age is associated with money.

Design and Methods: We collected a random sample of individuals and for each determined their age **(recorded in years)** and the amount of money (in dollars) in their accounts. Analysis of the data was done using **<u>linear regression</u>**.

For statistic b_1 :
95% C.I. = [0.05, 1.05]
p-value = 0.036

Results: We obtained a random sample of $n = 9$ subjects. There is a statistically significant association between age and money (p-value =0.036). **For every additional year in age, an individual's amount of money increases on average by an estimated of \$0.55** (95% C.I. = [\$0.05, \$1.05]).

Conclusions: We found that, as hypothesized, age is associated with money. In our sample age accounted for about half of the variability observed in money (R^2=0.49). We **<u>predict</u>** that a 50 year old will have \$45.1 (95% P.I. = [\$5.6, \$84.5]), whereas a 40 year old will have \$39.6 (95% P.I. = [\$0.8, \$78.4]).

Small Print: The analysis rests on the following assumptions:

- the observations are independently and identically distributed.
- the **response** variable, money, is normally distributed.
- Homoscedasticity of residuals or equal variance.
- the <u>relationship</u> between **response** and **predictor** variables is linear.

2.1.3 Least squares solution

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

$$\hat{b}_1 = r_{xy} s_y / s_x$$

In terms of summary statistics, the least squares line is:

$$\widehat{y} = \hat{b}_0 + \hat{b}_1 x \quad = \quad (\overline{y} - \hat{b}_1 \overline{x}) + \hat{b}_1 x \tag{2.24}$$

2.1.3 Least squares solution

$$\hat{b}_0 = \overline{y} - \hat{b}_1\overline{x}$$

$$\hat{b}_1 = r_{xy}s_y/s_x$$

In terms of summary statistics, the least squares line is:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x = (\overline{y} - \hat{b}_1\overline{x}) + \hat{b}_1 x \tag{2.24}$$

$$= \overline{y} + \hat{b}_1(x - \overline{x}) \tag{2.25}$$

2.1.3 Least squares solution

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

$$\hat{b}_1 = r_{xy} s_y / s_x$$

In terms of summary statistics, the least squares line is:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x = (\overline{y} - \hat{b}_1 \overline{x}) + \hat{b}_1 x \tag{2.24}$$

$$= \overline{y} + \hat{b}_1 (x - \overline{x}) \tag{2.25}$$

$$= \overline{y} + s_y r_{xy} \frac{x - \overline{x}}{s_x}, \tag{2.26}$$

2.1.3 Least squares solution

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

$$\hat{b}_1 = r_{xy} s_y / s_x$$

In terms of summary statistics, the least squares line is:

$$\widehat{y} = \hat{b}_0 + \hat{b}_1 x = (\overline{y} - \hat{b}_1 \overline{x}) + \hat{b}_1 x \tag{2.24}$$

$$= \overline{y} + \hat{b}_1 (x - \overline{x}) \tag{2.25}$$

$$= \overline{y} + s_y r_{xy} \frac{x - \overline{x}}{s_x}, \tag{2.26}$$

$$\frac{\widehat{y} - \overline{y}}{s_y} = r_{xy} \frac{x - \overline{x}}{s_x}. \tag{2.27}$$

2.1.3 Least squares solution

$$\hat{b}_0 = \overline{y} - \hat{b}_1 \overline{x}$$

$$\hat{b}_1 = r_{xy} s_y / s_x$$

In terms of summary statistics, the least squares line is:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x = (\overline{y} - \hat{b}_1 \overline{x}) + \hat{b}_1 x \tag{2.24}$$

$$= \overline{y} + \hat{b}_1 (x - \overline{x}) \tag{2.25}$$

$$= \overline{y} + s_y r_{xy} \frac{x - \overline{x}}{s_x}, \tag{2.26}$$

$$\frac{\hat{y} - \overline{y}}{s_y} = r_{xy} \frac{x - \overline{x}}{s_x}. \tag{2.27}$$

Equation (2.27) is the easiest way to remember the least squares line (read as "standardized y = correlation coefficient times standardized x"); from this, the intercept $\hat{b}_0$ and slope $\hat{b}_1$ can be obtained using simple algebra.

Chapter 2

- **Section 2.1** has the mathematics leading to the least squares line.

- **Section 2.2** introduces the simple linear regression model (prediction with one explanatory variable) that is formulated for a predictive equation. This is needed to quantify the variability of the coefficients of the best-fitting line, when different samples are taken from the population.

- **Section 2.5** has intervals for simple linear regression: the confidence interval for the slope of the least square line, confidence intervals for subpopulation means, and prediction intervals for a future or out-of-sample Y given x^*.

- **Section 2.6** has an explanation of Student t quantiles used in the interval estimates.