# STAT 135, Concepts of Statistics

Helmut Pitters

Linear regression

Department of Statistics
University of California, Berkeley

April 20, 2017

# Linear regression.

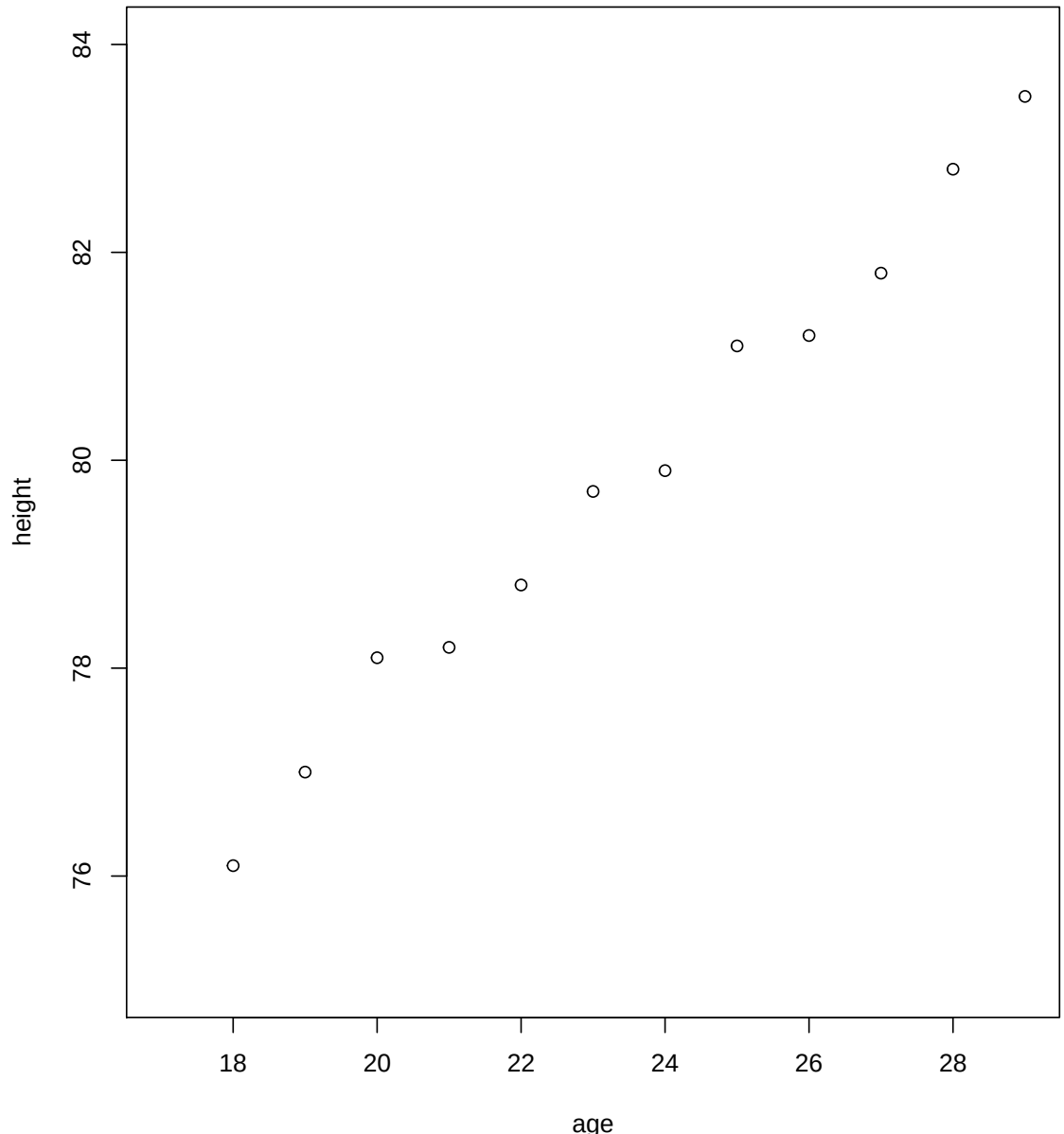
**Example (Growth of Kalama children)**

How (fast) do children grow?

In context of nutritional study in Egyptian village Kalama heights of $n = 161$ children [1] were recorded from their $18$th to $29$th month. Next figure shows scatterplot of average heights.

[1] Sampled randomly among all children in Kalama in their $18$th month.

Figure: Age (months) and height (cm) of 161 Kalama children.



# Linear regression.

## Example (Growth of Kalama children)

| age (months) | height (cm) |
|---|---|
| 18 | 76.1 |
| 19 | 77.0 |
| 20 | 78.1 |
| 21 | 78.2 |
| 22 | 78.8 |
| 23 | 79.7 |
| 24 | 79.9 |
| 25 | 81.1 |
| 26 | 81.2 |
| 27 | 81.8 |
| 28 | 82.8 |
| 29 | 83.5 |

Table: Age and height of Kalama 161 children.

# Linear regression.

**Example (Growth of Kalama children)**

Consider another child, of $25$ months say, (sampled randomly) from Kalama.
Question: How would you guess the child's height?

# Linear regression.

## Example (Growth of Kalama children)

Guessing height of $25$ month old child.
Heuristic:
Scatterplot displays linear pattern—correlation coefficient $r = 0.994$.

- Draw "the" line suggested by the scatterplot that comes "as close as possible" to the data points (*fitting a line*).
- Read off $y$-value at $x = 25$.

Challenges:

- What do we mean by "the" line?
- How do we find "the" line?

# Linear regression.

**Example (Growth of Kalama children)**

Clearly, no straight line can fit the data points

$$(x_i, y_i) = (x_i\text{th month}, \text{average age in } x_i\text{th month})$$

exactly. Some of the data points will be missed, which incurs an error.
However, a good candidate for "the line" among all straight lines should be the one that minimizes this error.

# Linear regression.

There are numerous ways to define the distance between a straight line and a point $(x_i, y_i)$.
Suppose line we are looking for is given by

$$f(x) = \beta_1 x + \beta_0,$$

i.e. has

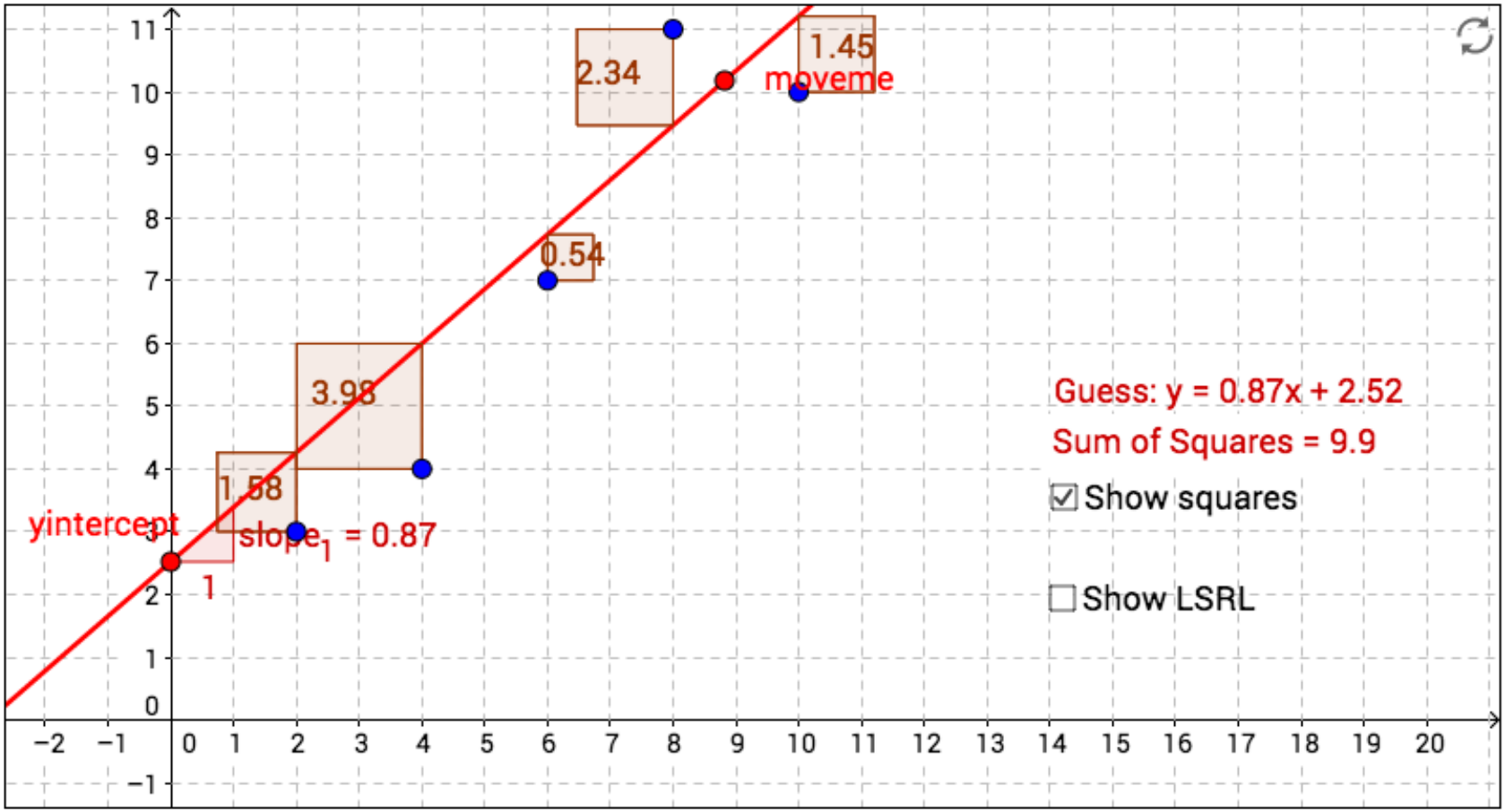
$$\begin{aligned} \text{slope} \quad & \beta_1 \\ \text{intercept} \quad & \beta_0. \end{aligned}$$

Agree to define distance, or error, between $(x_i, y_i)$ and $f(x)$ by squared distance

$$(y_i - f(x_i))^2 = (y_i - \beta_1 x_i - \beta_0)^2.$$

# Linear regression.

Figure: Finding the regression line, http://www.geogebra.org/m/105271.



# Linear regression.

Overall error between line and data points is sum of errors at each data point, i.e.

$$S(\beta_0, \beta_1) := \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2.$$

So, formally our task is to find $\hat{\beta}_0$, $\hat{\beta}_1$ such that

$$S(\beta_0, \beta_1) \text{ is minimised.}$$

Then

$$\hat{r}(x) = \hat{\beta}_1 x + \hat{\beta}_0$$

is the line we are looking for, the so-called *fitted line*.
This method of minimizing $S(\beta_0, \beta_1)$ is called the *method of least squares*.

# Linear regression.

## Theorem

*Slope and intercept of the regression line are given by*[2]

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r\sqrt{\frac{s_{yy}}{s_{xx}}}, \qquad \hat{\beta}_0 = -\hat{\beta}_1 \bar{x} + \bar{y}.$$

## Proof.

Exercise. Hint: Find roots of partial derivatives

$$\frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) = -2 \sum_{i=1}^n (y_i - \beta_1 x_i - \beta_0)$$

$$\frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) = -2 \sum_{i=1}^n x_i (y_i - \beta_1 x_i - \beta_0).$$

□

[2] Notice that $r = 0$ if and only if $\hat{\beta}_1 = 0$.

# Linear regression.

**Proof.**

$$S(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

For root of partial w.r.t. $\beta_0$ find

$$0 = \frac{\partial}{\partial \beta_0} S = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) \implies \beta_0 n = n\bar{y} - \beta_1 n \bar{x}.$$

Simple observation:

$$\sum_i x_i y_i - \sum_i \bar{x}\bar{y} = \sum_i (x_i - \bar{x}) y_i = \sum_i (x_i - \bar{x})(y_i - \bar{y}),$$

thus, choosing $y = x$, $\sum_i x_i^2 - \sum_i \bar{x}^2 = \sum_i (x_i - \bar{x})^2$. □

# Linear regression.

## Proof.

$$S(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

Now find root of partial w.r.t. $\beta_1$.

$0 = \frac{\partial}{\partial \beta_1} S = -2 \sum_i (y_i - \beta_0 - \beta_1 x_i) x_i$ and, replacing $\beta_0$ by $\hat{\beta}_0$

$$\begin{aligned}
0 &= \sum_i y_i x_i - \beta_0 n \bar{x} - \beta_1 \sum_i x_i^2 \\
&= \sum_i y_i x_i - (\bar{y} - \beta_1 \bar{x}) n \bar{x} - \beta_1 \sum_i x_i^2 \\
&= \sum_i x_i y_i - \sum_i \bar{x}\bar{y} - \beta_1 \left( \sum_i x_i^2 - \sum_i \bar{x}^2 \right) \\
&= \sum_i (x_i - \bar{x})(y_i - \bar{y}) - \beta_1 \sum_i (x_i - \bar{x})^2.
\end{aligned}$$

# Linear regression.

## Example (Growth of Kalama children)

Output of command **lm(formula = height∼age)** in R:

--

age <- 18:29

height <- c(76.1, 77.0, 78.1, 78.2, 78.8, 79.7, 79.9, 81.1, 81.2, 81.8, 82.8, 83.5)

Call:

lm(formula = height ∼ age)

Coefficients:

(Intercept) age

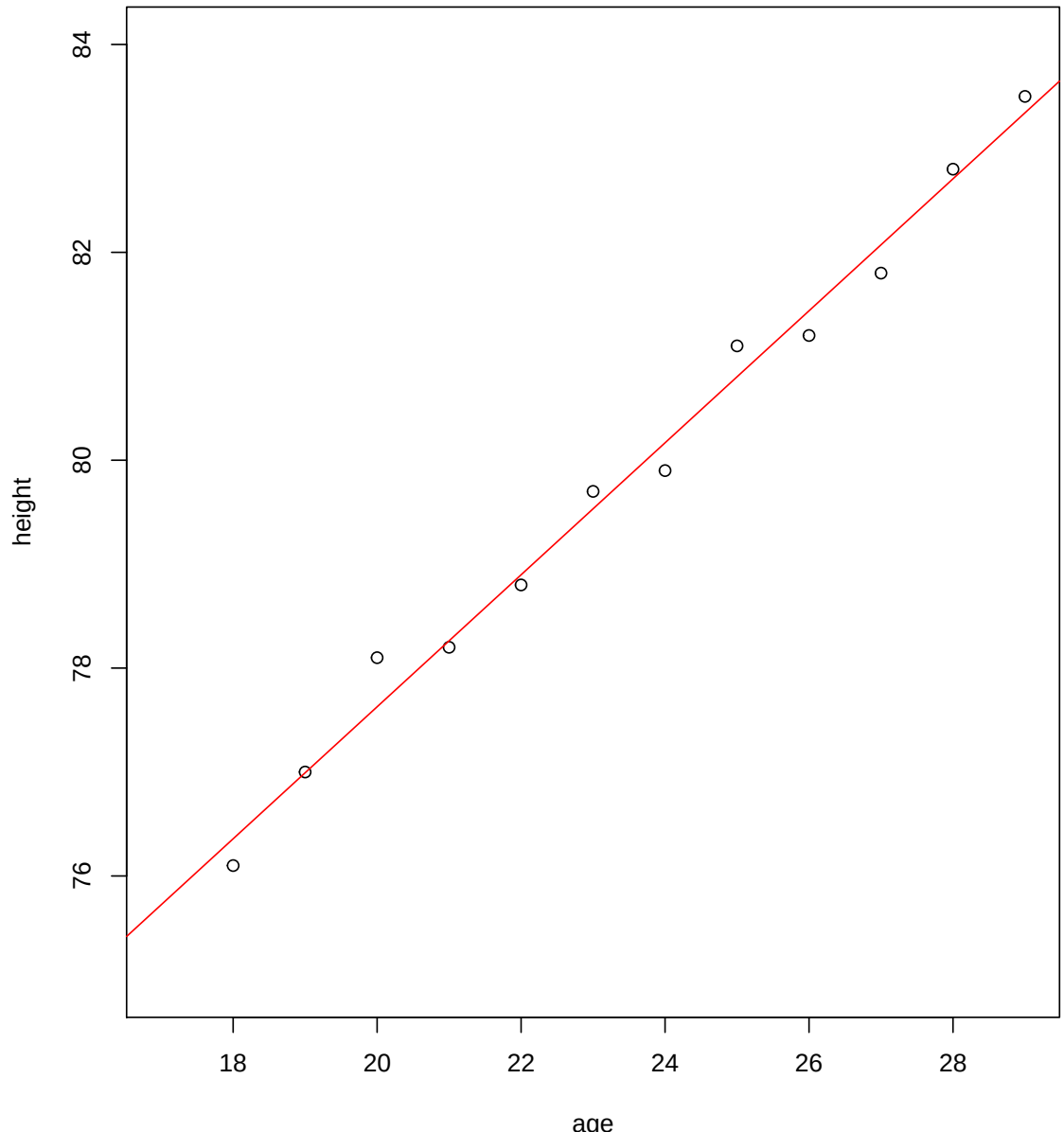
64.928 0.635

–

Consequently, the fitted line is given by

$$\hat{r}(x) = 0.635x + 64.93.$$

Figure: Age (months) and height (cm) of 161 Kalama children.



# Simple linear regression

# Simple linear regression.

So far, we have not made any assumption on the distribution underlying the data $(x_1, y_1), \ldots, (x_n, y_n)$.

Standard and most popular statistical model is *simple linear regression model*:

$$\text{(linearity)} \qquad y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{for } i \in \{1, \ldots, n\}$$

where $\epsilon_1, \ldots, \epsilon_n$ is an i.i.d. sequence of r.v.s such that

$$\mathbb{E}\epsilon_i = 0 \qquad \mathrm{Var}(\epsilon_i) = \sigma^2.$$

$x_i$ are assumed fixed, called *predictor* variables, and

$y_i$ are called *response* variables

$\beta_1$ can be interpreted as average change of $y_i$ if $x_i$ increases by $1$ unit

$\beta_0, \beta_1$ are called the regression coefficients

# Simple linear regression.

## Remark

- Want to stress that here we model the $x_i$ as being deterministic (think of a controlled experiment were the investigator can freely choose the $x_i$.). More general models of linear regression allow to model the $x_i$ as random.
- Notice that $\mathbb{E}y_i = \beta_0 + \beta_1 x_i, \quad \mathrm{Var}(y_i) = \sigma^2$.
  I.e. if we would repeatedly draw samples $(x_1, y_1), \ldots, (x_n, y_n)$ and average out the $y_i$ values for each $x_i$, we'd obtain the line $x \mapsto \beta_0 + \beta_1 x$ through the points $(x_i, \mathbb{E}y_i)$.
- Here "simple" refers to the $x_i$ being one-dimensional.
- Unkown parameters are: $\beta_0, \beta_1, \sigma^2$. Will use $\hat{\beta}_0, \hat{\beta}_1$ as estimators for $\beta_0, \beta_1$.

# Simple linear regression.

If we had taken another sample $(x_1, y_1), \ldots, (x_n, y_n)$ from the same model, i.e. such that

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \text{for } i \in \{1, \ldots, n\},$$

would have obtained different values $\hat{\beta}_0, \hat{\beta}_1$ for optimal intercept and slope of regression line by method of least squares.
Suppose we wanted to use fitted line

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

to make predictions.

$\hat{y}_i := \hat{r}(x_i)$ are the *predicted/fitted* values
$\hat{\epsilon}_i := y_i - \hat{r}(x_i)$ are the *residuals*.

How reliable are $\hat{\beta}_0, \hat{\beta}_1$ as estimators for $\beta_0, \beta_1$?

# Simple linear regression.

Recall estimators of intercept and slope given by least squares method:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = -\hat{\beta}_1 \bar{x} + \bar{y}.$$

Using $\sum_i (x_i - \bar{x}) = 0$ and $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, which implies $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{\epsilon}$, where $\bar{\epsilon} := \frac{1}{n} \sum_{i=1}^n \epsilon_i$, we obtain

$$\mathbb{E}\hat{\beta}_1 = \mathbb{E}\frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i - \beta_0 - \beta_1 \bar{x} - \bar{\epsilon})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1,$$

since $\mathbb{E}\epsilon_i = 0$, and therefore

$$\mathbb{E}\hat{\beta}_0 = \mathbb{E}[\bar{y} - \hat{\beta}_1 \bar{x}] = \beta_0.$$

# Simple linear regression.

We just proved

**Theorem**

*For simple linear regression, the estimators*

$$\hat{\beta}_0 \text{ and } \hat{\beta}_1$$

*for intercept and slope of the regression line are unbiased.*

# Simple linear regression.

Again, recall

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Thus, for the variance we find

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{\sum_i x_i^2 - n\bar{x}^2} = \frac{n\sigma^2}{n\sum_i x_i^2 - (\sum_i x_i)^2}.$$

# Simple linear regression.

**Theorem**

*For $\hat{\beta}_0, \hat{\beta}_1$ given by the method of least squares under simple linear regression we have*

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{n\sigma^2}{n\sum_i x_i^2 - (\sum_i x_i)^2},$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\sigma^2 \sum_i x_i}{n\sum_i x_i^2 - (\sum_i x_i)^2} = -\bar{x}\,\mathrm{Var}(\hat{\beta}_1),$$

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_i x_i^2}{n\sum_i x_i^2 - (\sum_i x_i)^2} = \frac{\sum x_i^2}{n}\,\mathrm{Var}(\hat{\beta}_1).$$

# Simple linear regression.

**Proof.**

Let's first compute (co)variances of $y_i, \bar{y}$ using the multilinearity of covariance:

$$\mathrm{Var}(y_i) = \sigma^2 \quad \mathrm{Cov}(y_i, y_j) = \delta_{ij}\sigma^2,$$

where $\delta_{ij}$ equals 1 if $i = j$ and 0 otherwise,

$$\mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n} \quad \mathrm{Cov}(y_i, \bar{y}) = \frac{1}{n}\sum_{j=1}^{n} \mathrm{Cov}(y_i, y_j) = \frac{\sigma^2}{n}$$

and

$$\mathrm{Cov}(\bar{y}, \hat{\beta}_1) = \frac{\sum (x_i - \bar{x})\,\mathrm{Cov}(\bar{y}, y_i)}{\sum (x_i - \bar{x})^2} = 0, \text{ and therefore}$$
$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \mathrm{Cov}(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1) = -\bar{x}\,\mathrm{Var}(\hat{\beta}_1).$$

□

**Proof.**

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

$$\begin{aligned}
\operatorname{Var}(\hat{\beta}_0) &= \operatorname{Var}(\bar{y}) + \operatorname{Var}(\hat{\beta}_1 \bar{x}) - 2\operatorname{Cov}(\bar{y}, \hat{\beta}_1 \bar{x}) \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \operatorname{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sigma^2 (\sum (x_i - \bar{x})^2 + n\bar{x}^2)}{n \sum (x_i - \bar{x})^2} = \frac{\sum x_i^2}{n} \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sum x_i^2}{n} \operatorname{Var}(\hat{\beta}_1).
\end{aligned}$$

□

# Simple linear regression.

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_i x_i^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}, \mathrm{Var}(\hat{\beta}_1) \quad = \frac{n\sigma^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}.$$

In practice, parameter $\sigma^2$ is usually unknown. However, since

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \hat{r}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

residuals

$$\hat{\epsilon}_i = y_i - \hat{r}(x_i)$$

should capture the information about $\sigma^2$ contained in the sample. We'll see that

$$s^2 := \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

is unbiased estimator for $\sigma^2$. The expression $RSS := \sum_{i=1}^{n} \hat{\epsilon}_i^2$ is called the *residual sum of squares (RSS)*.

# Simple linear regression.

Consequently, replacing $\sigma^2$ by $s^2$, we estimate the variances of $\hat{\beta}_0, \hat{\beta}_1$ by

$$s^2_{\hat{\beta}_0} = \frac{s^2 \sum_i x_i^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}$$

$$s^2_{\hat{\beta}_1} = \frac{n s^2}{n \sum_i x_i^2 - (\sum_i x_i)^2}.$$

# Simple linear regression.

Recall simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $(\epsilon_i)$ are i.i.d. r.v.s with $\mathbb{E}\epsilon_i = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$.
If we assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (which is the case in numerous settings), then responses $y_i$ are normally distributed. In particular,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = -\hat{\beta}_1 \bar{x} + \bar{y}.$$

are linear combinations of normally distributed r.v.s, hence normal.
One can show that

$$\frac{\hat{\beta}_0 - \beta_0}{s_{\hat{\beta}_0}} \text{ and } \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{n-2}.$$

Consequently, one can construct confidence intervals and hypothesis tests for $\beta_0, \beta_1$.

# Simple linear regression.

**Example (Growth of Kalama children)**

R command: **summary(lm(formula = height∼age))**.

# Simple linear regression.

For calculations, the formulas

$$\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_j - \sum x_i \sum x_j y_j}{n \sum x_i^2 - (\sum x_i)^2}$$
$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_j}{n \sum x_i^2 - (\sum x_i)^2}$$

can be convenient.

**Proof.**

Recall $\sum_i (x_i - \bar{x}) = \sum x_i^2 - n\bar{x}^2 = \sum x_i^2 - n^{-1}(\sum x_i)^2$.

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum (x_i - \bar{x})^2},$$

now multiply both numerator & denominator by $n$. □

# Simple linear regression.

$$\hat{\beta}_0 = \frac{\sum x_i^2 \sum y_j - \sum x_i \sum x_j y_j}{n \sum x_i^2 - (\sum x_i)^2}$$

$$\hat{\beta}_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_j}{n \sum x_i^2 - (\sum x_i)^2}$$

**Proof.**

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{\sum y_i \sum (x_j - \bar{x})^2}{n \sum x_i^2 - (\sum x_i)^2} - \frac{n\bar{x} \sum x_i y_i - \bar{x} \sum x_i \sum y_j}{n \sum x_i^2 - (\sum x_i)^2},$$

and

$$\begin{aligned} \sum y_i \left( \sum x_j^2 - n\bar{x} \right) + \sum y_i \bar{x} \sum x_j &= \sum y_i \left( \sum x_j^2 - n\bar{x}^2 + n\bar{x}^2 \right) \\ &= \sum y_i \sum x_j^2. \end{aligned}$$

□

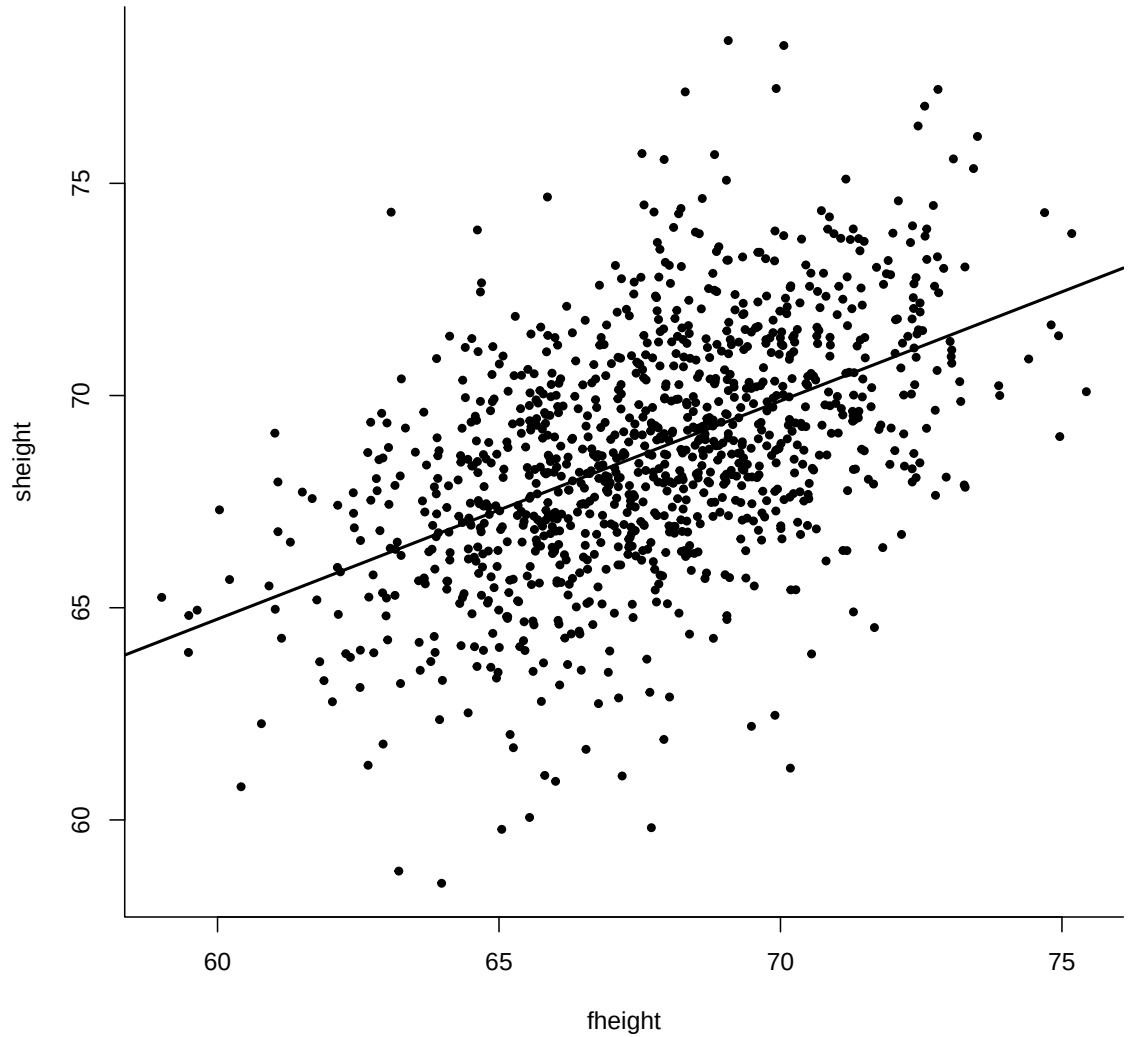
# Regression to the mean.

> I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning. When I had finished my enthusiastic speech, one of the most seasoned instructors in the audience raised his hand and made his own short speech, which began by conceding that positive reinforcement might be good for the birds, but went on to deny that it was optimal for flight cadets. He said, "On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don't tell us that reinforcement works and punishment does not, because the opposite is the case."

Daniel Kahnemann, Nobel laureate in Economic Sciences

# Galton's study of heights of fathers and sons.

Figure: Height (inch) of fathers and their sons.



# Galton's study of heights of fathers and sons.

Children of larger than average parents tend to be smaller than their parents. Likewise, children of smaller than average parents tend to be larger than their parents.

Why?

# Regression to the mean.

> This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them. I immediately arranged a demonstration in which each participant tossed two coins at a target behind his back, without any feedback. We measured the distances from the target and could see that those who had done best the first time had mostly deteriorated on their second try, and vice versa. But I knew that this demonstration would not undo the effects of lifelong exposure to a perverse contingency.
>
> Daniel Kahnemann, Nobel laureate in Economic Sciences
> (from Kahnemann's biography at nobelprize.org)

# Linear regression.

A great number of models can be cast as linear via appropriate data transformations.

E.g. consider a model of exponential growth/decay for a population

$$P_t = P_0 e^{m(t-t_0)},$$

where

$$P_t = \text{ size of population at time } t$$
$$m = \text{ growth rate.}$$

Question: How to estimate growth rate from given data $(x_i, P_{x_i})$?

Idea: Apply log

$$\begin{aligned}\log P_t &= \log P_0 + m(t - t_0) = \log P_0 - mt_0 + mt \\ &= \beta_0 + \beta_1 t.\end{aligned}$$

Introducing random errors $(\epsilon_t)$ the model becomes

$$\log P_t = \beta_0 + \beta_1 t + \epsilon_t$$

and can be treated as a simple linear regression model.

# Simple linear regression.

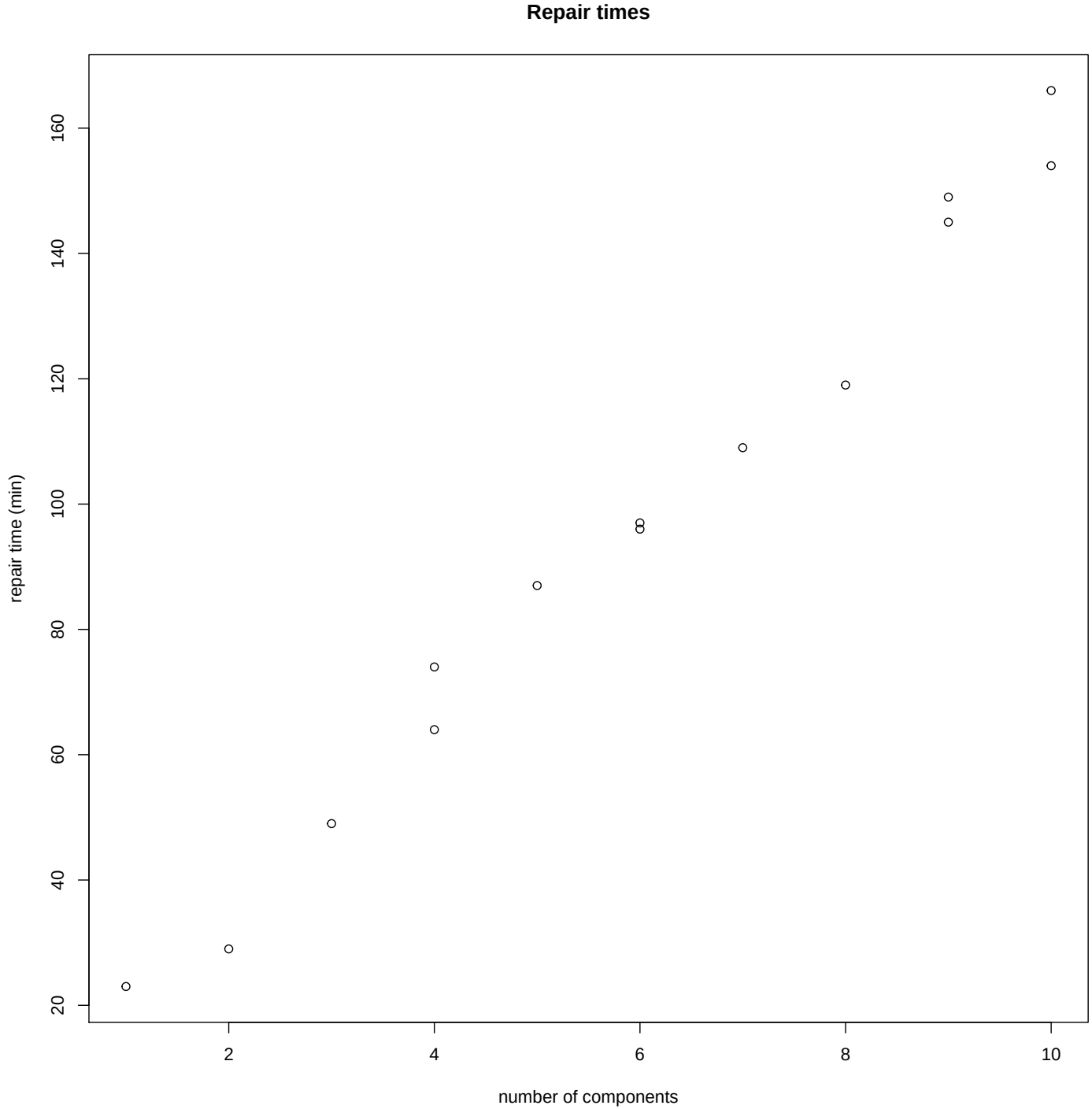
## Example (Repair times, 1)

Company repairs certain devices with small number of electronic components. Important for company's allocation of resources is to understand how lengths of service calls, $(y_i)$, depend on number of electronic components, $(x_i)$.

$x_i =$ number of electrical components in device for $i$th service

$y_i =$ length of $i$th service call.

Next figure shows plot of sample of size $n = 14$.


Repair times
repair time (min)
20
40
60
80
100
120
140
160
2
4
6
8
10
number of components

# Simple linear regression.

## Example (Repair times, 1)

From data we find

$$\bar{x} = 6, \qquad \bar{y} = 97.21, \qquad r = 0.996.$$

High correlation coefficient confirms linear association we see in scatter plot.
Seems appropriate to model relationship between number of components and repair time by simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $(\epsilon_i)$ are i.i.d. $\mathbb{E}\epsilon_i = 0$, $\mathrm{Var}(\epsilon_i) = 1$.

# Simple linear regression.

## Example (Repair times, 1)

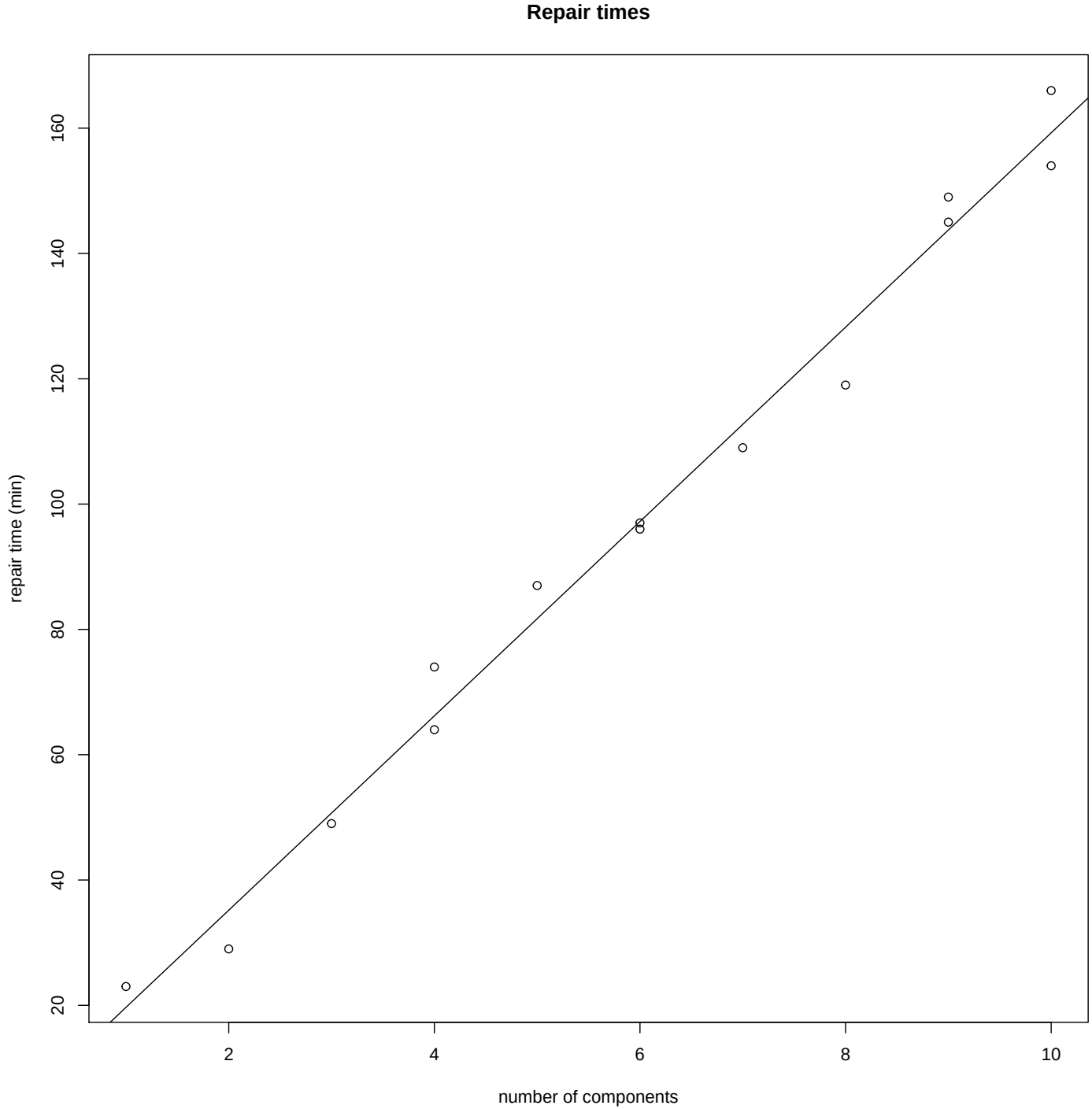
For least squares estimates obtain

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \hat{x})(y_i - \hat{y})}{\sum_i (x_i - \bar{x})^2} = 15.51,$$
$$\hat{\beta}_0 = -\beta_1 \bar{x} + \bar{y} = 4.16,$$

and thus the regression line

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = 4.16m + 15.51mx.$$

Let's check this result informally by looking at the plot of the regression line.


Repair times
repair time (min)
20
40
60
80
100
120
140
160
2
4
6
8
10
number of components

# Simple linear regression.

## Example (Repair times, 1)

regression line

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = 4.16m + 15.51mx.$$

Interpretation:

$\hat{\beta}_0 = 4.16m$ : time needed to set up repair service independent of number of components

$\hat{\beta}_1 = 15.5m$ : avg. increase in service time for each additional component.

However, one should be careful with interpretations. In particular the value $\hat{r}(0) = 4.16m$ is obtained by interpolating in a region where there are no observations (extrapolation)!

# Simple linear regression.

**Model deficiencies.** In addition to examining correlation coefficient $r$ and t statistics one should always consult a graphical plot of the data to see whether linear fit is actually meaningful.

To this end analysis of residuals $(\hat{\epsilon}_i)$ is often essential.

We now present four rather different data sets given by Anscombe (1973), all with same $r$ and t statistic. However, only one of them shows linear association.

R: plots of Anscombe's data sets.

# Simple linear regression.

**Detecting model deficiencies.** To detect wether assumptions of linear model might be violated, often residuals

$$\hat{\epsilon}_i = y_i - \hat{r}(x_i)$$

are examined graphically.
E.g. residuals $(\hat{\epsilon}_i)$ can be plotted against

- predictors: $(x_i, \hat{\epsilon}_i)$
- fitted values: $(\hat{y}_i, \hat{\epsilon}_i)$
- order in which observations occurred: $(i, \hat{\epsilon}_i)$.

Plots that do not appear symmetric about[3] $y = 0$ or show a distinct pattern of variation suggest that (some) assumptions of linear model are violated.

[3] A consequence of $\mathbb{E}\epsilon_i = 0$.

# Simple linear regression.

Assumption that residuals $(\epsilon_i)$ are normally distributed can be checked via probability plots.

**Example (Repair times, 1)**

R: Analysis of residuals.
Analysis does not contradict our assumptions of a linear model.
Whether random disturbances $(\epsilon_i)$ can be assumed to be normally distributed is not so clear.

# Simple linear regression.

Since residuals have different variances, often *standardized residuals*

$$\hat{\epsilon}_i^{\mathrm{s}} := \frac{\hat{\epsilon}_i}{\text{estimated SE}(\hat{\epsilon}_i)}$$

are examined instead.
Here estimated standard error of $\hat{\epsilon}_i$ is given by[4]

$$s\sqrt{1 - \frac{1}{n} - \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2}}.$$

[4] We don't prove this.

# Simple linear regression.

## Example (Repair times, 1)

One way to assess how good predictors $(x_i)$ explain responses $(y_i)$ in linear model is to check hypotheses

$$H_0 \colon \hat{\beta}_1 = 0 \qquad \text{vs} \qquad H_A \colon \hat{\beta}_1 \neq 0.$$

Null hypothesis can be interpreted as there being almost no linear association between $(x_i)$ and $(y_i)$[5].
Obtain for t-statistic $(df = 12)$

$$\frac{\hat{\beta}_1 - \beta_1}{\mathsf{SE}(\hat{\beta}_1)} = 30.71, \qquad (\text{since } \mathsf{SE}(\hat{\beta}_1) = 0.505)$$

and p-value $8.92 \times 10^{-13}$. Thus $H_0$ can be safely rejected.

[5] Recall $\hat{\beta}_1 = r s_x / s_y$.

# Simple linear regression.

**Example (Repair times, 1)**

A $(1-\alpha) = 95\%$ confidence interval for $\beta_1$ is given by

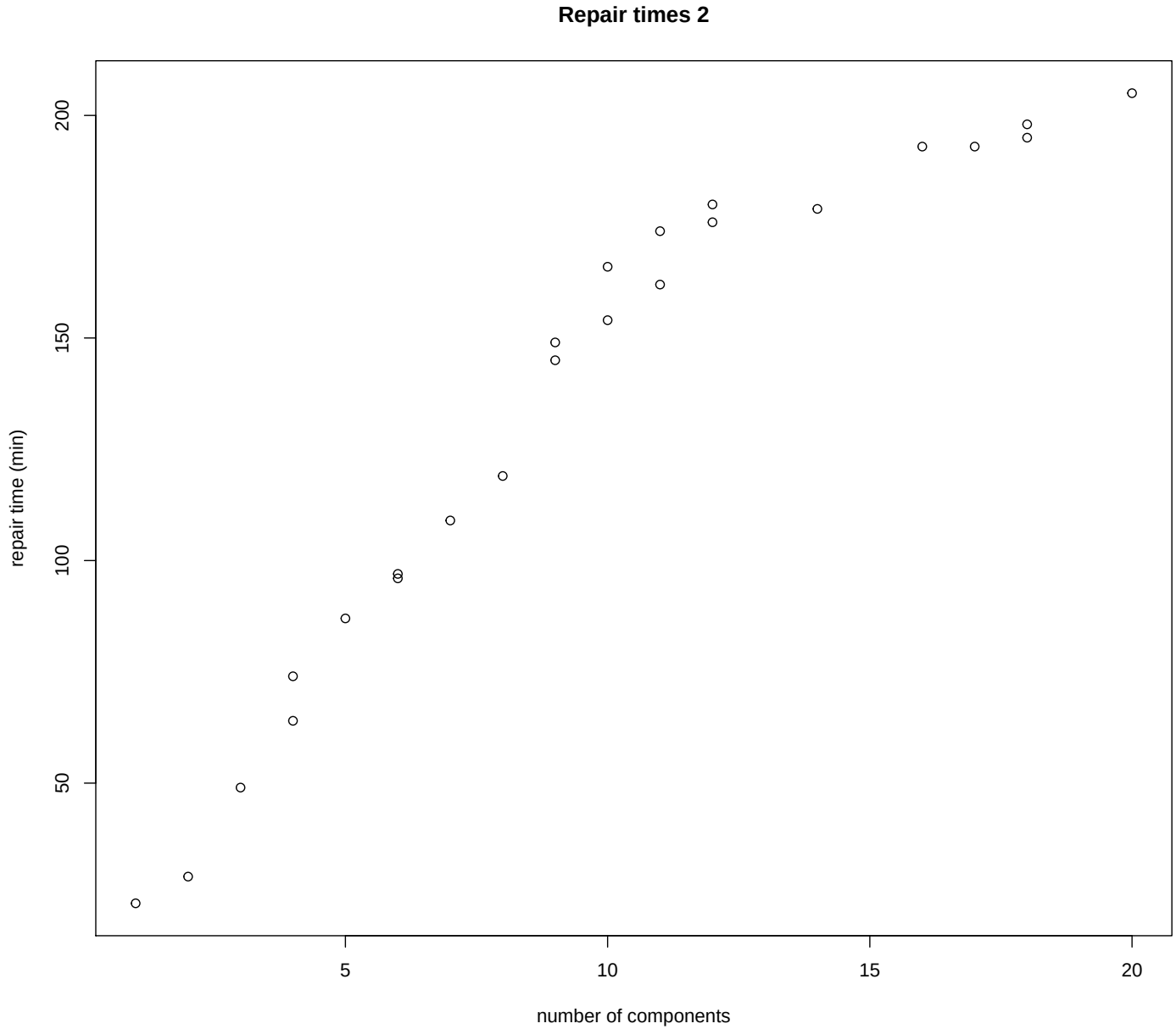
$$\hat{\beta}_1 \pm \mathsf{SE}(\hat{\beta}_1) t_{12}(\frac{\alpha}{2}) = 15.51m \pm 0.505m \times 2.18 = [14.41m, 16.61m].$$

Interpretation: If we observe a large number of repair times, in $90\%$ of cases the avg. increase in repair time per component is between $14.41m$ and $16.61m$

# Simple linear regression.

**Example (Repair times, 2)**

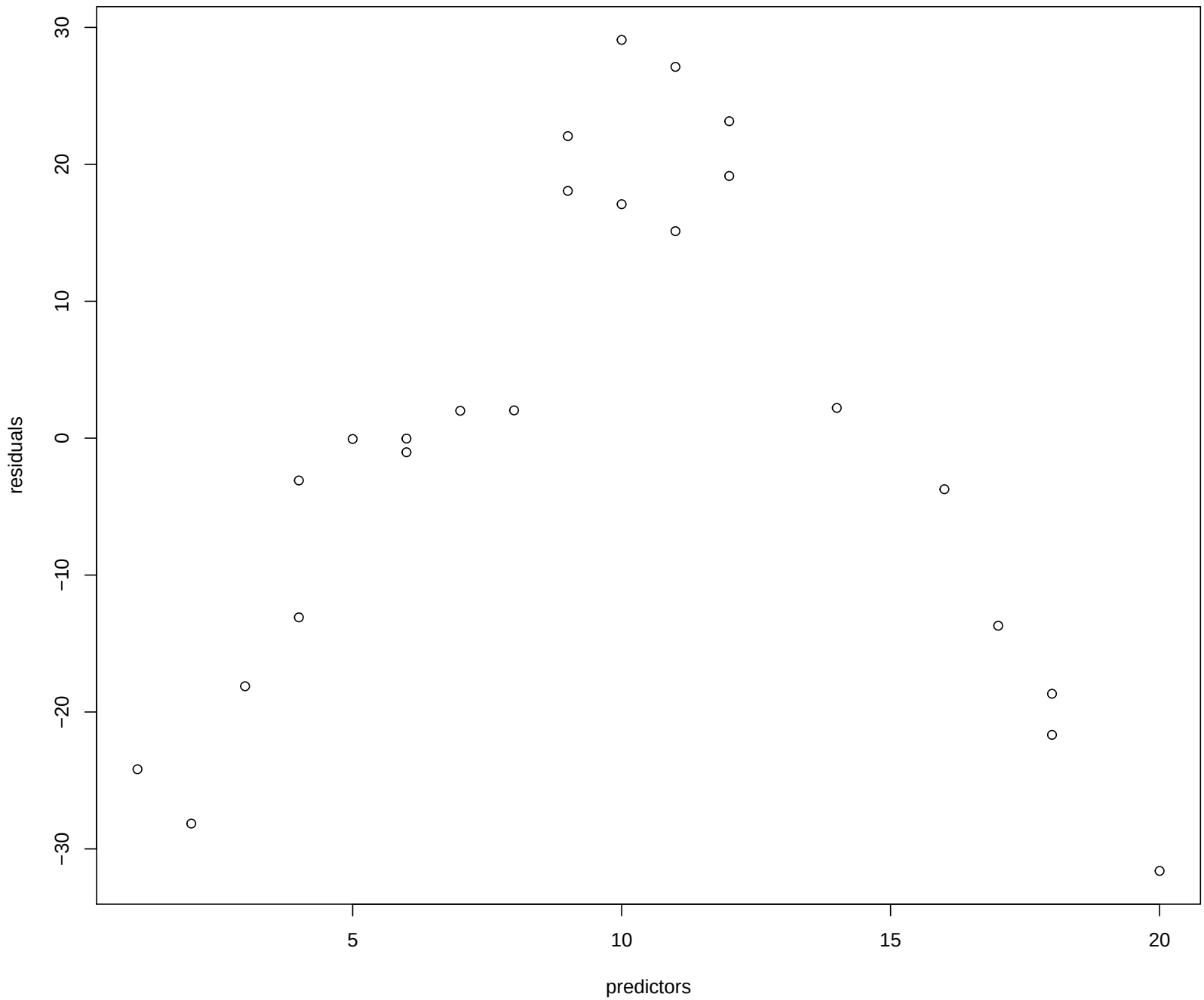
In another sampling period (same sampling method, same company) $10$ further observations were taken. Next figure shows their scatter plot.


Repair times 2
repair time (min)
50
100
150
200
5
10
15
20
number of components

# Simple linear regression.

**Example (Repair times, 2)**

There seems to be a break point at $x \approx 12, 13$ components, where slope declines.
Suggests an effect of time efficiency when device for repair has $12$ or more components.
R: analyzing residuals.


Analysis of residuals 2: residuals vs predictors
residuals
predictors
30
20
10
0
−10
−20
−30
5
10
15
20

# Simple linear regression.

**Example (Repair times, 2)**

Linear regression does not seem to be appropriate to proceed with statistical analysis.
Investigator may want to study reasons behind observed efficiency.
This can lead to observation of another important factor influencing repair time (besides # of components) –> multiple regression?!

# Multiple linear regression

# Multiple linear regression.

Saw before that in simple linear regression response $y_i$ is modeled to depend on one predictor $x_i$ in a linear fashion:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i. \tag{1}$$

Often don't expect that response can be well predicted by one predictor only. E.g. don't expect income of person to only depend on years of education, but also on socioeconomic status of parents, work experience, etc.
Therefore, allow $y_i$ to depend on several predictors $x_{i,1}, \ldots, x_{i,p-1}$:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i,$$

where $(\epsilon_i)$ are i.i.d., $\mathbb{E}\epsilon_i = 0$, $\mathrm{Var}(\epsilon_i) = \sigma^2$, and $\beta_0, \ldots, \beta_p$ are unknown parameters.
This is the so-called *multiple linear regression model*.

# Multiple linear regression.

There are two particularly important statistical regimes depending on the ratio of the number of $p$ and the number of observations:

- $\frac{p}{n} \to 0$ as $n \to \infty$. Classical regime (that we consider).
- $\frac{p}{n} \to C > 0$ as $n \to \infty$. A regime that emerged in more recent years (high dimensional data, e.g. data on health status of patient). Harder to study, less tools available, but more interesting.

# Simple linear regression.

To study multiple linear regression, turns out that notions from linear algebra are very well suited. We collect some of these notions and introduce some notation.

**Notions needed from linear algebra.** Consider

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n.$$

Often consider $x$ as $n \times 1$ matrix, i.e. an element of $\mathbb{R}^{n \times 1}$.

Euclidean norm: $\|x\| := \sqrt{\sum_{i=1}^n x_i^2}$.