Correlation and linear regression

Manuela Zucknick

Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
manuela.zucknick@medisin.uio.no

MF9130E – Introductory Course in Statistics
09-05-2023

Outline

Aalen chapter 11.1-11.3, Kirkwood and Sterne chapter 10

- Learn how to model a relationship between two variables: **correlation** & (simple) **linear regression**
- **Assumptions** to be checked when using linear regression analysis
- Example R code
- (Appendix: some **theory** behind regression)

Outline for this afternoon

12.45-14.15: Regression analysis I: Simple regression, correlation.
Literature: Aalen chap. 11.1-11.3, K&S chap. 10

14.30-15.15: R exercise for regression I.

15.15-16.00: Discussion of the R exercise for regression I in class.

What is regression about?

- Measuring several quantities.
- Aim: detecting the *association* between them.
- Regression is a statistical method for analysing association.
- It is closely related to correlation.

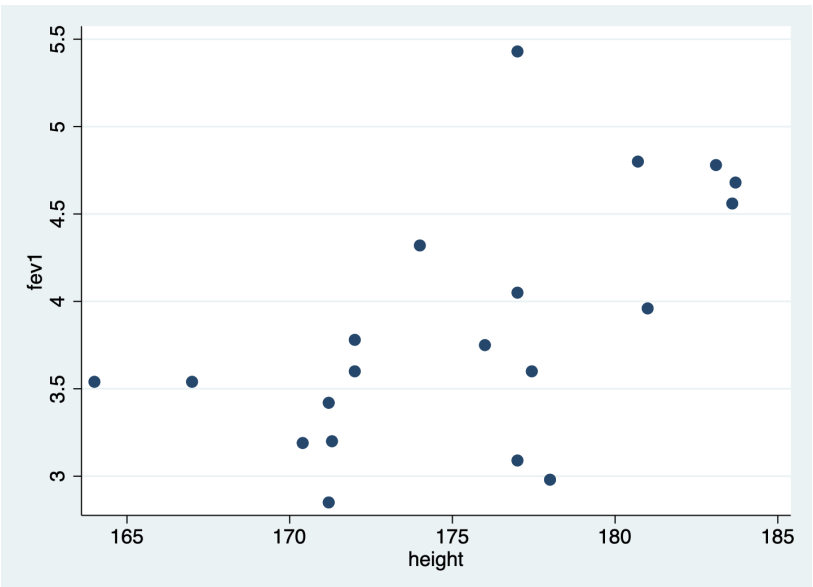
Correlation: How close are the points to a straight line?

- Correlation is always between -1 and $+1$.
- Correlation $+1$ means that the points lie on a straight line with positive slope (-1 correlation means negative slope).
- Correlation 0 means no association.

Underlying example: FEV1 and height

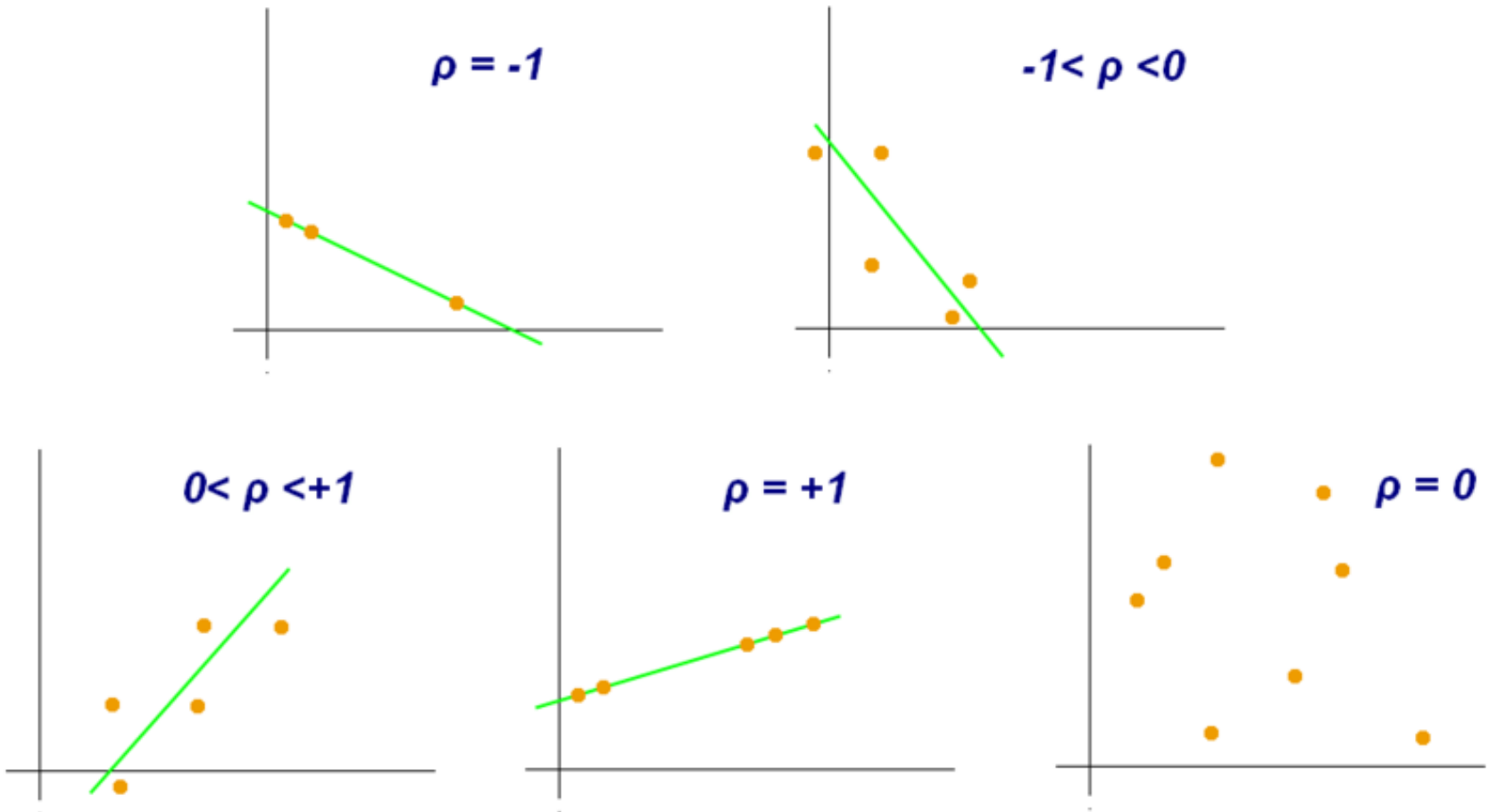
- We want to examine the association between two continuous variables,
- and we use a toy example: lung function (FEV1) and height for 20 male medical students.

height	fev1
174	4.32
180.7	4.8
183.7	4.68
177	5.43
177	3.09
172	3.78
176	3.75
177	4.05
164	3.54
178	2.98
167	3.54
171.2	3.42
177.44	3.6
171.3	3.2
183.6	4.56
183.1	4.78
172	3.6
181	3.96
170.4	3.19
171.2	2.85



Pearson's coefficient of correlation: r

- This is a measure of *linear trend* associated with two variables X and Y,
- $-1 \leq r \leq 1$,

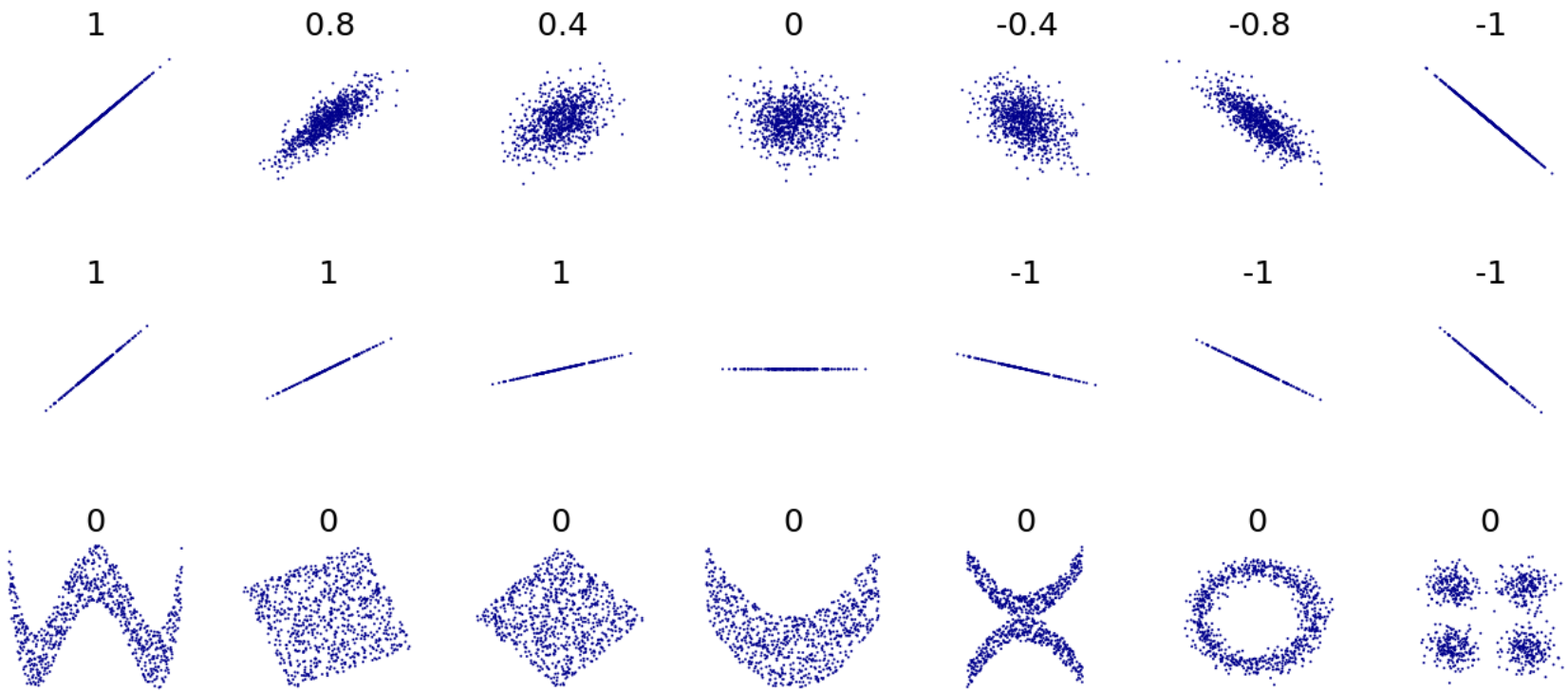


Definition of r

- Variables X and Y,
- Outcomes $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$,

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}} \tag{1}$$

Scatter plots: examples of r



- Scatter plots are useful to explore your data,
- R: `plot()`.

Correlations in R

- R: `cor()`.
- Note that the result will be `NA` (missing value), if there are any missing values in the data.
  In this case, use option `use = "complete.obs"`.

```
> FEV1 <- read.csv("FEV1.csv")
>
> #Pearson's correlation:
> cor(FEV1$fev1, FEV1$height)
[1] 0.5810765
>
> #If there might be missing data, use argument use="complete.obs":
> cor(FEV1$fev1, FEV1$height, use="complete.obs")
[1] 0.5810765
```

Testing for a correlation

Aim

We are often interested in testing whether a sample correlation r is large enough to indicate a nonzero population correlation.

This corresponds to testing the following hypothesis:

H_0: the true correlation equals 0,

H_1: the true correlation is different from 0.

- Test statistic: $\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2}$

- **Assumption:** variables have to be normal, correlation has to be linear. Always check the scatterplot!

Testing for a correlation in R

```
> #Correlation with test against H0: r = 0:
> cor.test(FEV1$fev1, FEV1$height)

        Pearson's product-moment correlation

data:  FEV1$fev1 and FEV1$height
t = 3.0292, df = 18, p-value = 0.007212
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1865162 0.8142280
sample estimates:
      cor
0.5810765

>
> #Alternative using the formula (~):
> cor.test(~ fev1 + height, data=FEV1)
```

- Use function `cor.test()` for a statistical test of the null hypothesis that the correlation is zero.
- Also provides a confidence interval.

Pairwise correlation of many variables

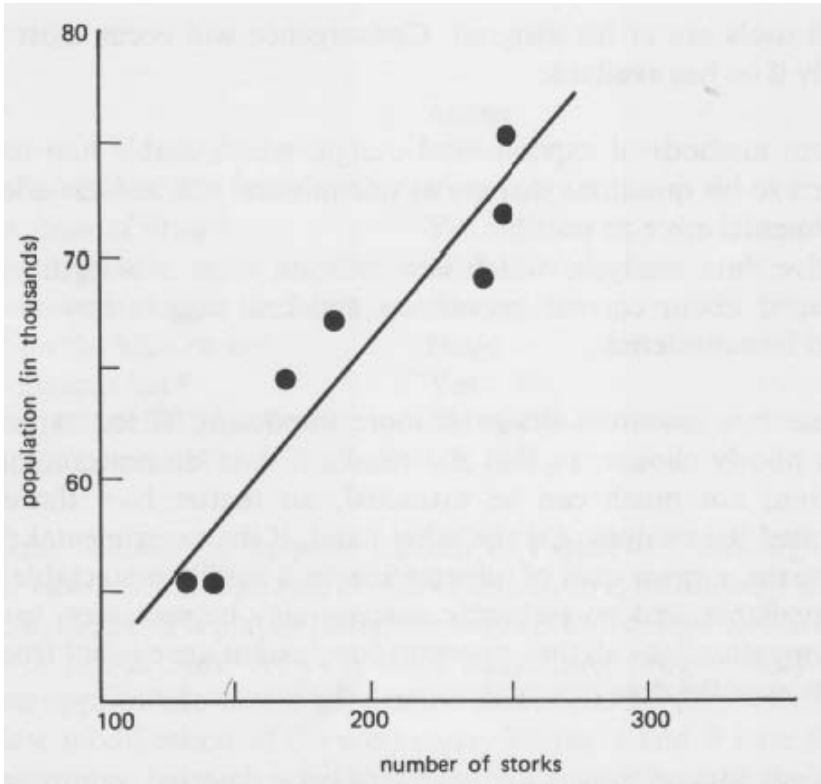
R will calculate a matrix of all pair-wise correlations if we provide the data in form of a matrix or dataframe.

Take for instance the data set with PEF measurements:

```
> cors <- cor(PEF[,c("height", "weight", "pefmean",
+                    "pefsit1", "pefsit2", "pefsit3")],
+             use="complete.obs")
> print(cors, digits=2)
        height weight pefmean pefsit1 pefsit2 pefsit3
height    1.00   0.83    0.69    0.68    0.68    0.67
weight    0.83   1.00    0.70    0.71    0.70    0.67
pefmean   0.69   0.70    1.00    0.98    0.99    0.98
pefsit1   0.68   0.71    0.98    1.00    0.97    0.96
pefsit2   0.68   0.70    0.99    0.97    1.00    0.98
pefsit3   0.67   0.67    0.98    0.96    0.98    1.00
```

A few warnings: confounders

- A large r can in some situations be due to a third variable (confounder), and does not necessarily represent a causal relation.
- Example: in the figure, correlation between human birth rate and number of storks



Non-linear trends

- A small r does not imply that there is no trend, only that there is no linear trend,
- r is therefore not suitable in the following situation:

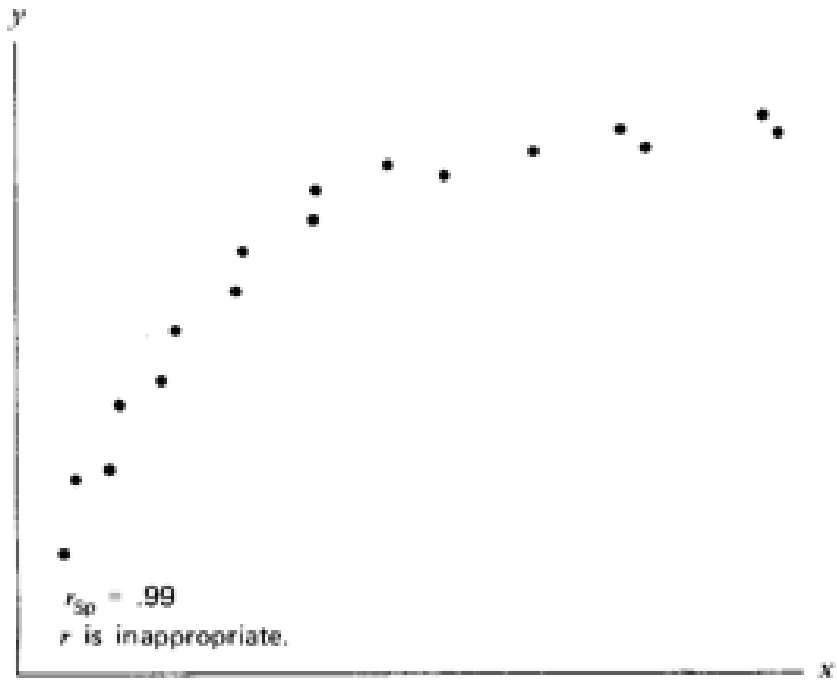
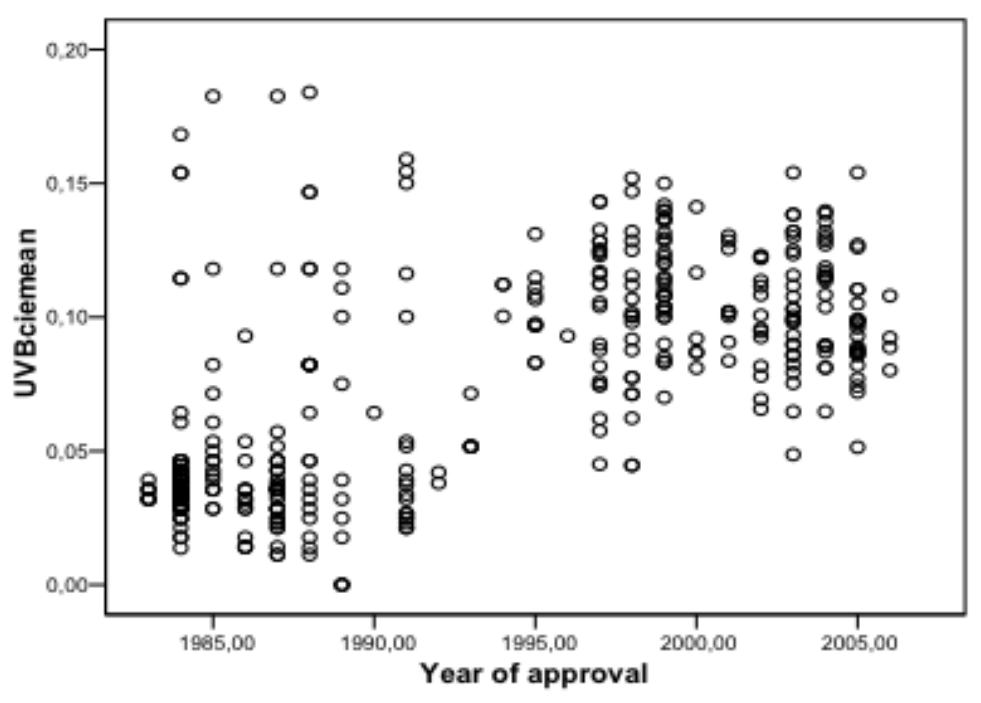


Figure 15.5 r_{Sp} is a measure of any monotone relationship.

Clustered scatter plots

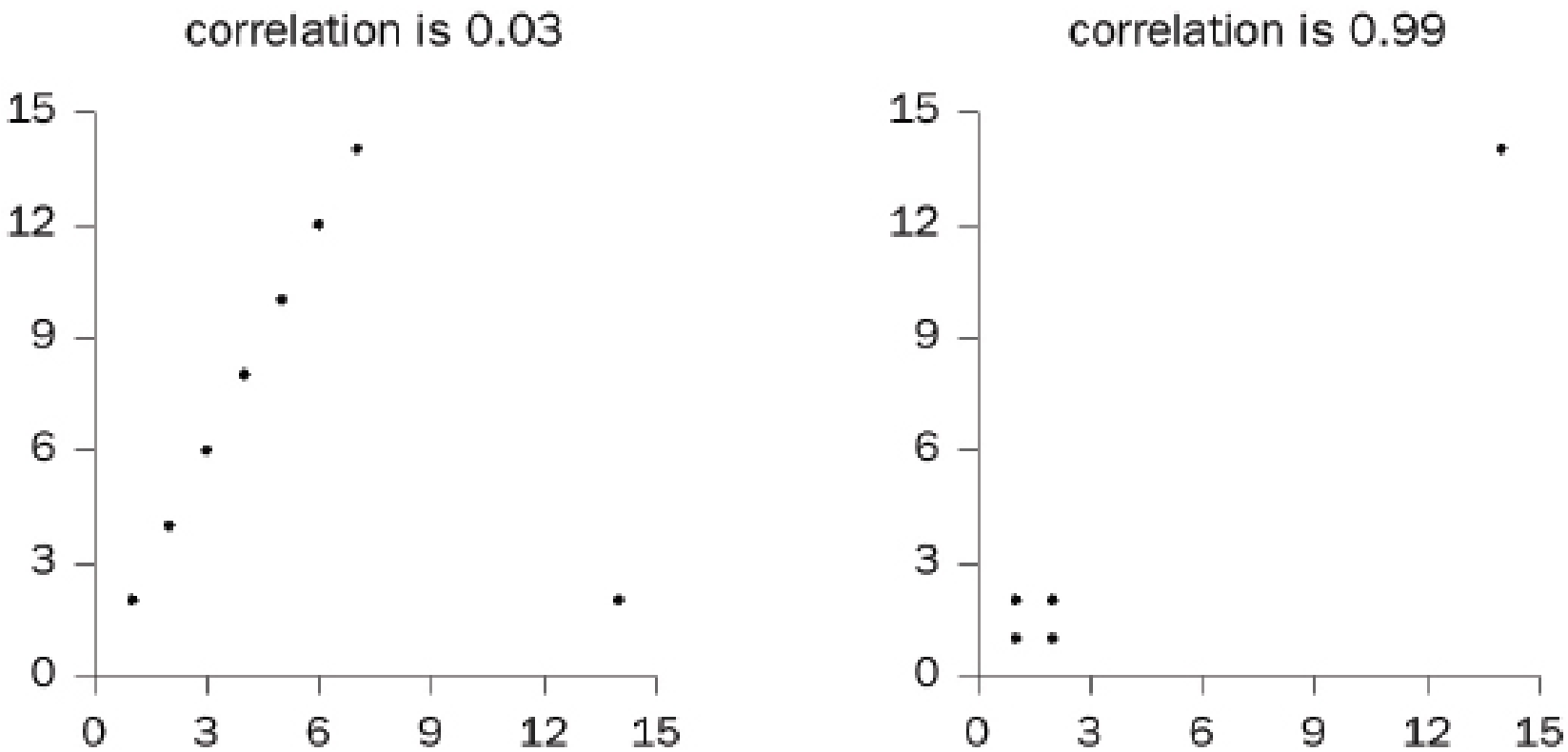
- The correlation coefficient is not suitable if the scatter plot consists of separate clusters,



- 1983-2005: $r = 0.65$
- 1983-1992: $r = 0.16$
- 1993-2005: $r = 0.07$

Outliers

- Some observations can have too big impact on the correlation,

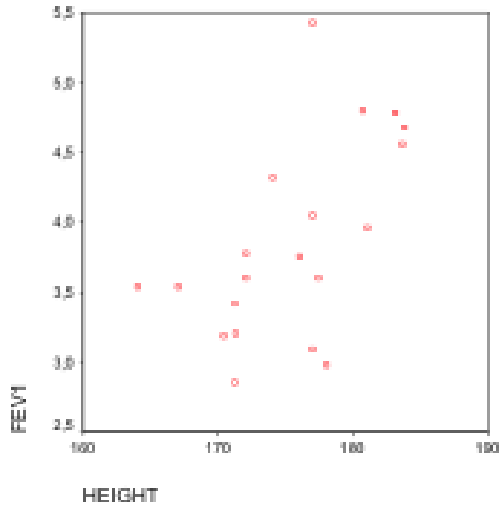


- It can be a good idea to remove such outliers (e.g. for testing sensitivity of the regression results to individual data points).

FEV1 vs. height: Linear regression

Table 11.1. FEV1 and height for 20 male medical students

Height (cm)	FEV1 (litre)	Height (cm)	FEV1 (litre)
174.0	4.32	167.0	3.54
180.7	4.80	171.2	3.42
183.7	4.68	177.4	3.60
177.0	5.43	171.3	3.20
177.0	3.09	183.6	4.56
172.0	3.78	183.1	4.78
176.0	3.75	172.0	3.60
177.0	4.05	181.0	3.96
164.0	3.54	170.4	3.19
178.0	2.98	171.2	2.85



- The correlation $r = 0.58$ and the scatter plot indicate that there is a relation between FEV1 and height,
- We can quantify this relation with a *linear regression* analysis.

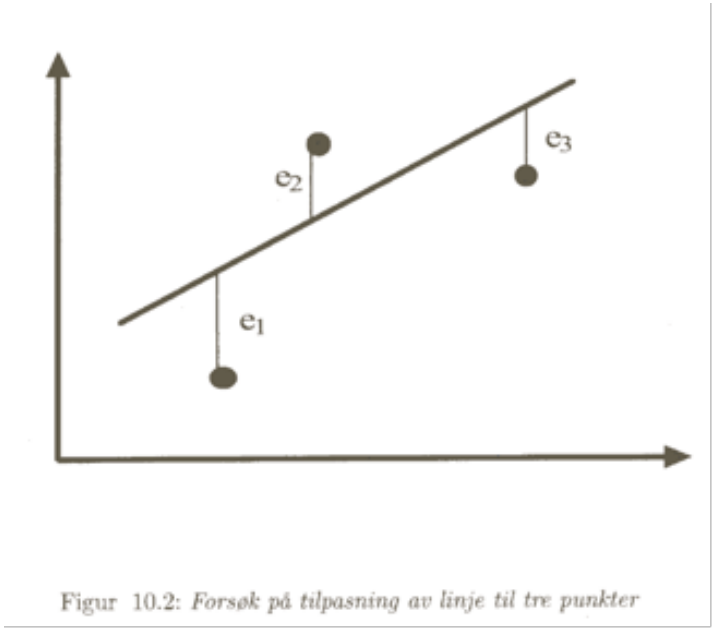
A line that represents the relation among FEV1 and height

- General formula for a straight line:

$$y = a + bx$$

- x and y are the coordinates of any point on the line;
- a and b are the line *parameters*:
  - a is where the line crosses the y-axis (**intercept**),
  - b is the **slope** of the line, i.e., its inclination with respect to the x-axis;
- a and b are *the only two parameters* that we need to estimate in order to completely define the regression line.

How do we find a and b?



Figur 10.2: *Forsøk på tilpasning av linje til tre punkter*

- There are no straight lines that go through all the measurements,
- The vertical distances from the measurements and the line are called residuals, and named: e_1, e_2, e_3.
- We will choose a and b such that the line minimizes *the sum of squared residuals*:

$$e_1^2 + e_2^2 + e_3^2.$$

Method of least squares

General method:

- Compute each residual with respect to the line $y = a + bx$,
- Compute the sum of squares of these numbers,
- Choose a and b that give the smallest sum of squares.

- See the appendix for some technical details.

Example: FEV1 vs. height

- In our toy example we are interested in how FEV1 changes with height
- In R, we use the command `lm()` (for "linear model") for all linear regression models.
- Note the `formula` notation for regressing y on x: `y ~ x`

```
> fit <- lm(fev1 ~ height, data=FEV1)
> fit

Call:
lm(formula = fev1 ~ height, data = FEV1)

Coefficients:
(Intercept)       height
   -9.18373      0.07435
```

- Estimated regression line:

$$\text{FEV1} \approx -9.18 + 0.07 \cdot \text{height}$$

Example: FEV1 vs. height

- As with many other analyses in R, we can use generic methods like `summary()` and `plot()` for more results:

```
> summary(fit)

Call:
lm(formula = fev1 ~ height, data = FEV1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.07065 -0.32340  0.03458  0.31794  1.45370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.18373    4.30672  -2.132  0.04700 *
height       0.07435    0.02454   3.029  0.00721 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5893 on 18 degrees of freedom
Multiple R-squared:  0.3376,    Adjusted R-squared:  0.3009
F-statistic: 9.176 on 1 and 18 DF,  p-value: 0.007212
```

Example: FEV1 vs. height

- $\hat{a} = -9.184$, $\hat{b} = 0.074$
- SE$(\hat{a}) = 4.307$, SE$(\hat{b}) = 0.025$,

- $t = 3.029$ gives $p = 0.007$, so $H_0 : b = 0$ is rejected at the 5% level.

Example: FEV1 vs. height

- Use `confint()` to calculate confidence intervals for the regression coefficients:

```
> confint(fit)
                    2.5 %     97.5 %
(Intercept) -18.23180580 -0.1356563
height        0.02278394  0.1259170
```

- 95% confidence interval for b: $(0.023, 0.126)$

Explained variance r^2 (also called R^2)

- We have that $0 \leq r^2 \leq 1$,
- and (as obvious from notation) r^2 is the square of the linear correlation r.
- If r^2 is *large*, it means that the observations are close to the regression line,
- If r^2 is *small*, it means that the observations are not so close to the regression line,
- Interpretation: r^2 quantifies the *proportion of variation* in the data that is *explained* by the fitted linear regression model.

- Recall that $r^2 = 0.338$ in the FEV1-example,
- so in words we can say that there is 34% explained variation in this example.

Residuals

- The deviations of the observed outcomes from the regression line are called *residuals*,
- Residuals are computed as

$$
\begin{aligned}
e_1 &= y_1 - (\hat{a} + \hat{b}x_1) \\
e_2 &= y_2 - (\hat{a} + \hat{b}x_2) \\
&\vdots \\
e_n &= y_n - (\hat{a} + \hat{b}x_n)
\end{aligned}
$$

- A *standardized residual* is the residual divided by the *empirical standard deviation*:

$$
\frac{e_i}{\sqrt{\frac{1}{n-1}\sum_{j=1}^{n} e_j^2}}
$$

Fitted values

We can use the model to compute the individual predictions based on their height, i.e.

$$0.074 \cdot \text{height} - 9.184.$$

- This equation can be used to predict values for $\mathtt{y}$ (FEV1) for a certain value of $\mathtt{x}$ (height), e.g.
- for someone with a height of 1.80m:

$$0.074 \cdot 180 - 9.184 = 4.136.$$

Conditions for linear regression

The residuals shall be

- Approximately independent.
- Not be systematically related to any independent variable or 'fitted value'.
- Their variance should be approximately constant (and not depend on the size of the fitted values).

- Be normally distributed around 0 (only needed for inference),
- ... and in this case, most standardized residuals lie between approx. -2 and $+2$.

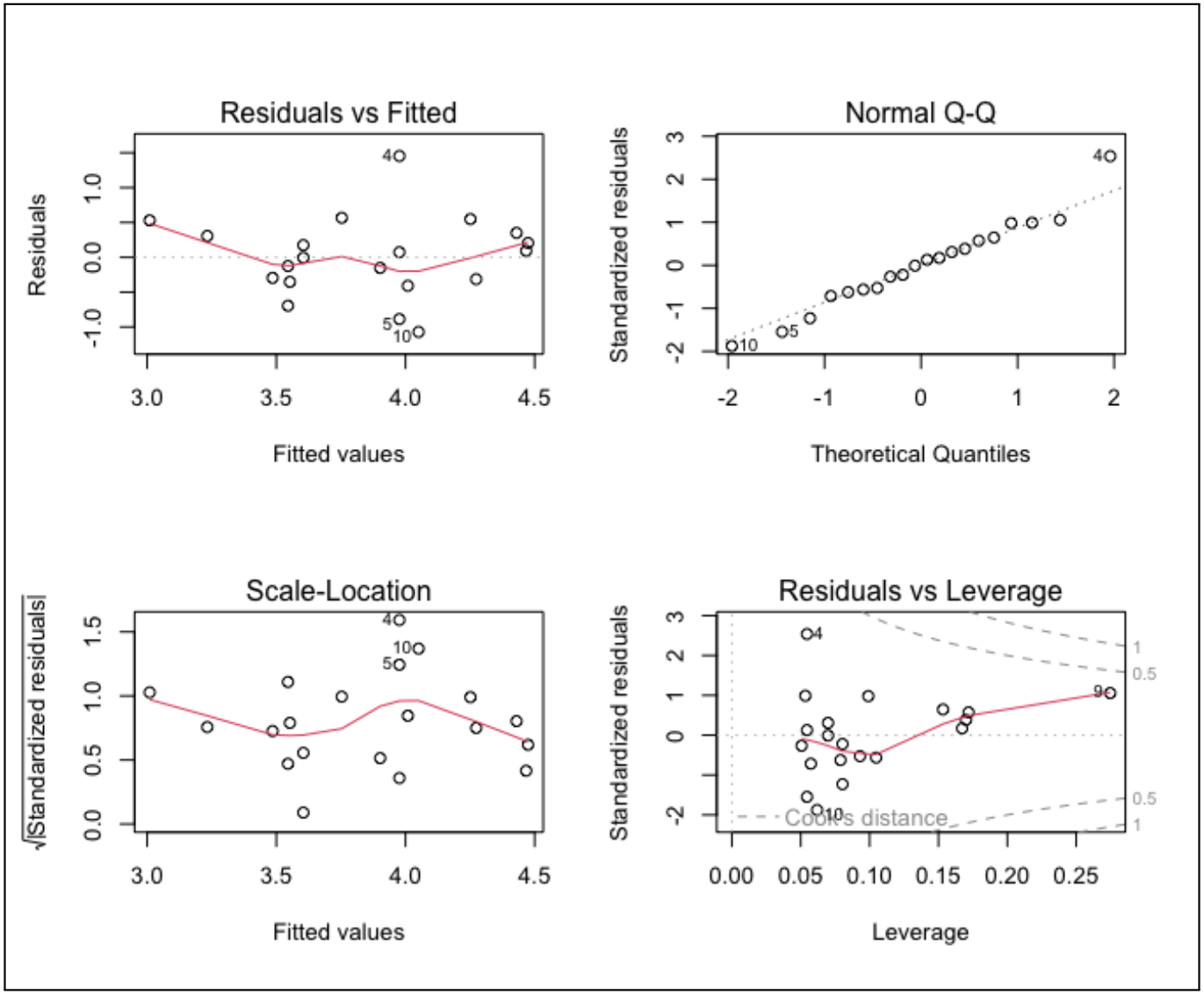
Using R to check the conditions for linear regression

- Normality plots for residuals, e.g. Normal Q-Q plots
  (top-right plot on next slide)
- Plot of (standardized) residuals versus fitted values
  (top-left and bottom-left plots on next slide)
- Plot of residuals versus covariates
  (slide after next)

- Examine large residuals and potential influence points with respect to their leverage
  (see tomorrow)

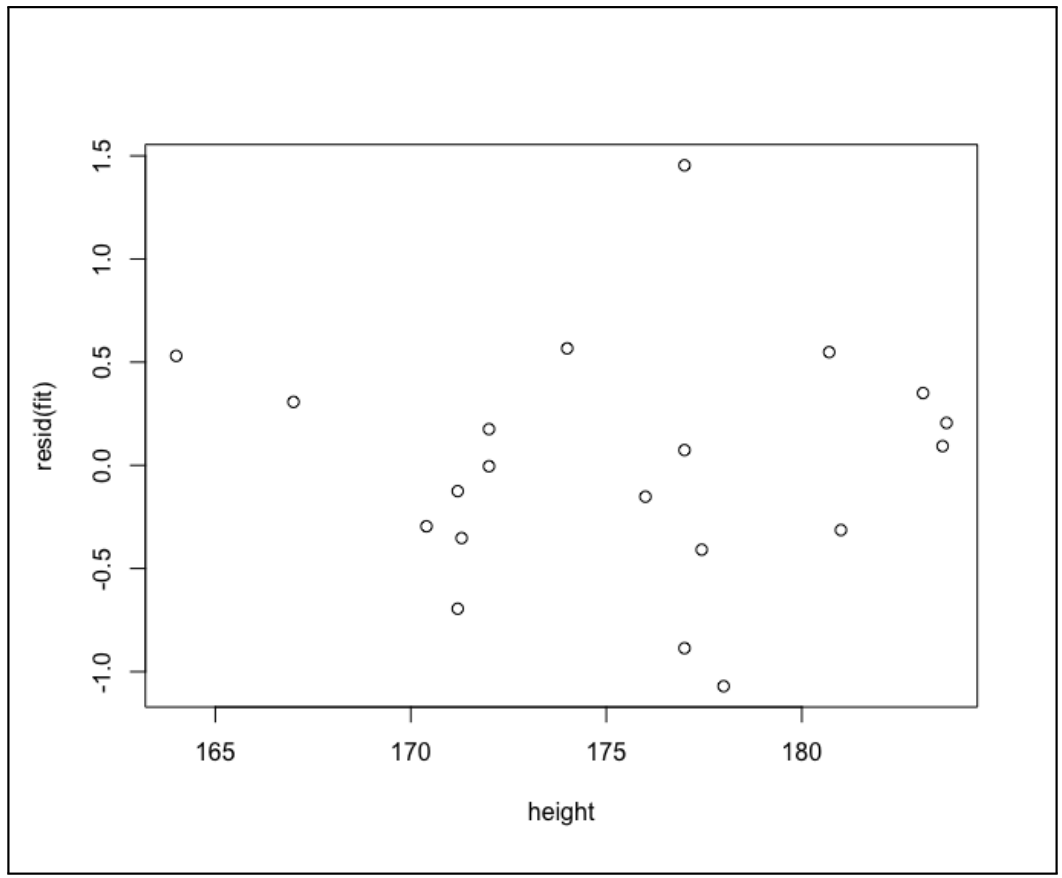
Example: FEV1 vs. height

```
> par(mfrow = c(2,2)) #arrange the following 4 plots 2-by-2
> plot(fit)
```



Plot of residuals versus covariates

```
> plot(resid(fit) ~ height, data=FEV1)
```



- This and the plots of residuals vs fitted values all indicate that the residuals are not systematically related to the fitted values/ covariates and are homoscedastic.

Summary

Key words

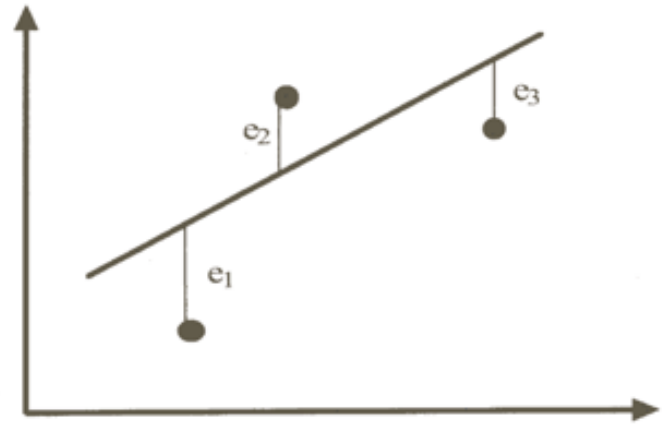
- Correlation
- Linear regression
- Residuals, conditions for linear regression

Notation

- r, r^2/ R^2
- a, b; (S_x, S_y)
- $e_1, \ldots, e_n$

Appendix: Some technical details

Minimizing the sum of squares



Figur 10.2: *Forsøk på tilpasning av linje til tre punkter*

Line with best fit is found by minimizing:

$$(y_1 - (a + bx_1))^2 + \cdots + (y_n - (a + bx_n))^2$$

Some calculus shows that $y = \hat{a} + \hat{b}x$ yields the best fitting line if

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

$$\hat{a} = \overline{y} - \hat{b}\overline{x}.$$

The estimated regression line is related to correlation!

Note that

$$\hat{\boldsymbol{b}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\boldsymbol{S_y}}{\boldsymbol{S_x}} \cdot \boldsymbol{r},$$

where

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{x})^2 \sum_{i=1}^{N}(y_i - \overline{y})^2}}$$

and

$$S_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \quad \text{and} \quad S_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}.$$

$\rightarrow$ S_x and S_y are *empirical standard deviations* of the variables x and y, respectively

95%-confidence interval for b

- Standard error of the regression coefficient estimator $\hat{b}$:

$$\mathsf{SE}(\hat{b}) := \frac{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))^2}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \tag{2}$$

- 95%-confidence interval for b:

$$\left(\hat{b} - c \cdot \mathsf{SE}(\hat{b}), \hat{b} + c \cdot \mathsf{SE}(\hat{b})\right),$$

where c is the 2.5% percentage point in the Student-t distribution with $n - 2$ degrees of freedom.

Testing the hypothesis $H_0 : b = 0$ vs. $H_A : b \neq 0$

- Test statistic:

$$T = \frac{\hat{b}}{\mathsf{SE}(\hat{b})} \tag{3}$$

- Under H_0 (i.e., if H_0 is true), T is distributed as a t-Student with $n - 2$ degrees of freedom
- This means that, given an observed test statistic T_0, the p-value equals

$$p = 2P(t_{n-2} \geq |T_0|). \tag{4}$$