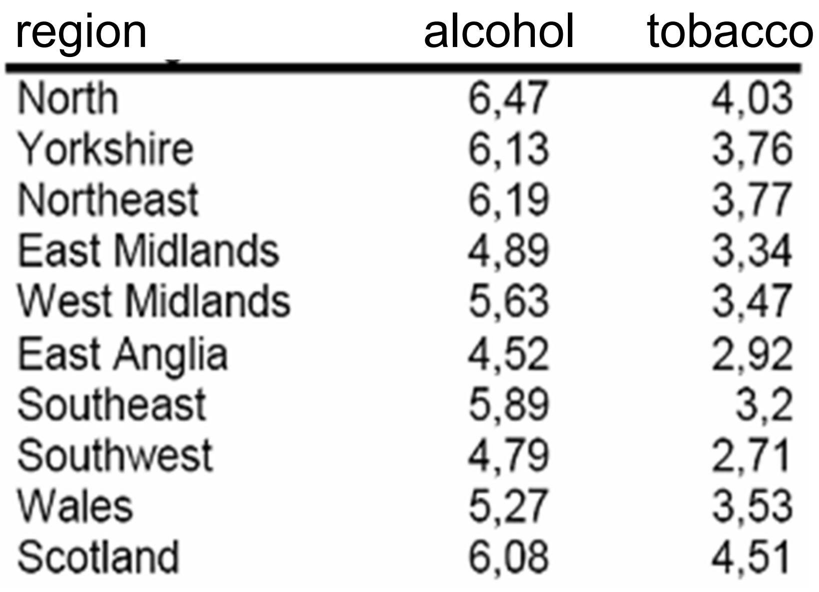
# Biostatistics week 10

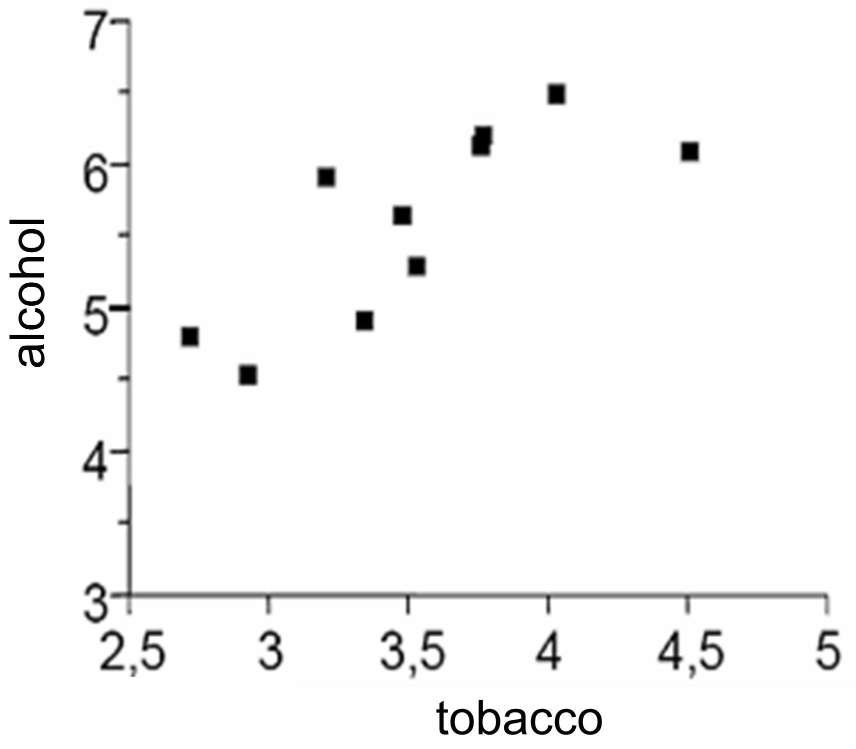
- Correlation and Rank correlation
- OLS estimates for coefficients in linear regression
- Coefficient of Determination $R^2$: unadjusted or adjusted
- Global F test: Compare model to Intercept model
- Compare nested model: F-test via R anova function
- Model and Variable selection with some warnings
- Linear regression with factor variables
  - interaction between a factor and a continuous predictor

# Is there an association between 2 variables?

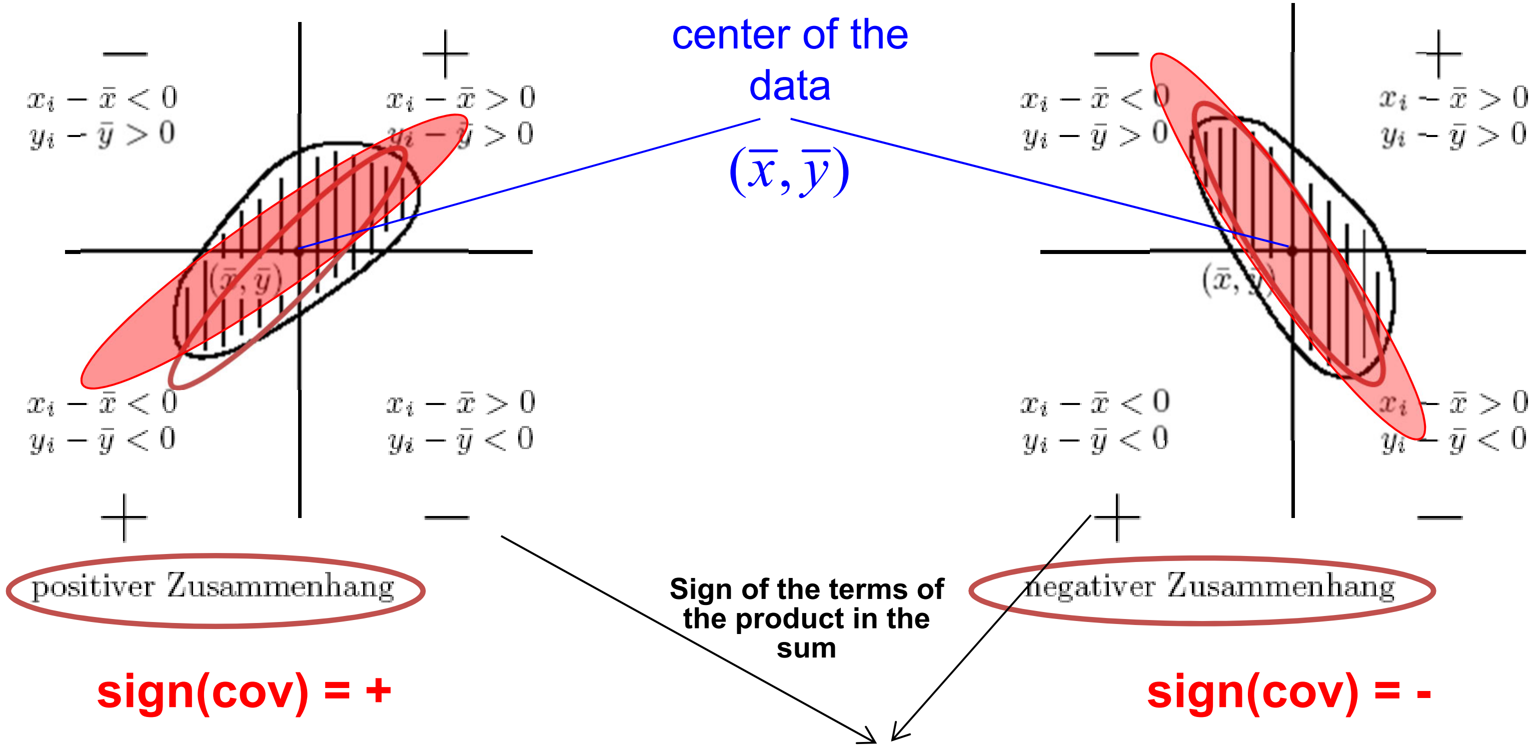
Example: observational study conducted in the UK:
Weekly expenses for alcohol and tobacco



| region | alcohol | tobacco |
|---|---|---|
| North | 6,47 | 4,03 |
| Yorkshire | 6,13 | 3,76 |
| Northeast | 6,19 | 3,77 |
| East Midlands | 4,89 | 3,34 |
| West Midlands | 5,63 | 3,47 |
| East Anglia | 4,52 | 2,92 |
| Southeast | 5,89 | 3,2 |
| Southwest | 4,79 | 2,71 |
| Wales | 5,27 | 3,53 |
| Scotland | 6,08 | 4,51 |



# Covariance determines the sign of a linear association



center of the data

$(\bar{x}, \bar{y})$

positiver Zusammenhang

Sign of the terms of the product in the sum

negativer Zusammenhang

**sign(cov) = +**

**sign(cov) = -**

$$\mathrm{cov}_{XY} = \frac{1}{n} \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

=> In R: cov(x,y)

# Covariance and the «standardized» correlation

Definition of the covariance:

$$\mathrm{cov}_{XY} = \frac{1}{n} \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

In mathematical statistics the covariance is often used. However, since the covariance depends on the scale of the variable (e.g. cm or m) the covariance is hardly used in data analysis.

The correlation is the better measure to quantify the strength of a linear relations, since the correlation is independent of the scale in which the variable was measured and ranges between +1 and -1.

covariance: cov(a*x,b*y) = a*b*cov(x,y)

correlation: cor(a*x,b*y) = cor(x,y)

# Pearson correlation coefficient

The Pearson correlation quantifies the strength and direction of a linear association between two variables x and y often observed at the same observation unit (e.g. height and weight of a person).

In a scatterplot y vs x or x vs y that means how closly the points scatter around a imagined straight line.

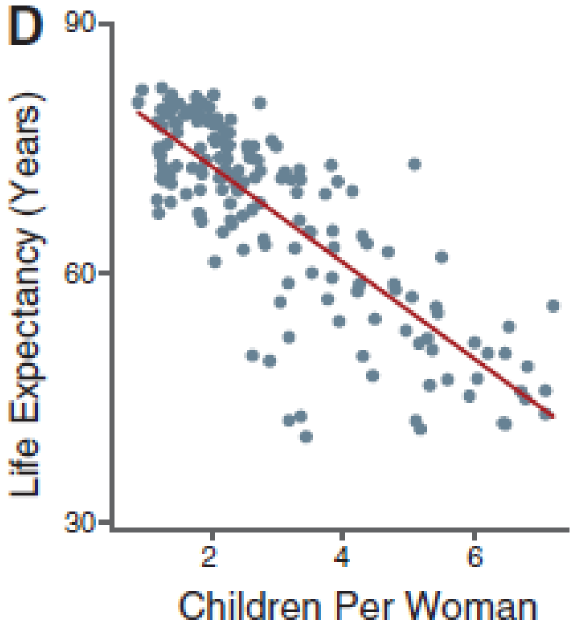
Definition:

$$r_{XY} = cor(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \cdot \sqrt{\sum (Y_i - \bar{Y})^2}} = \frac{\mathrm{cov(X, Y)}}{sd(X) \cdot sd(Y)}$$

=> In R: cor(x,y)

# Correlation – a word of warning

WHO has collected 357 variables (e.g. life expectancy and #children per woman) for 202 countries.

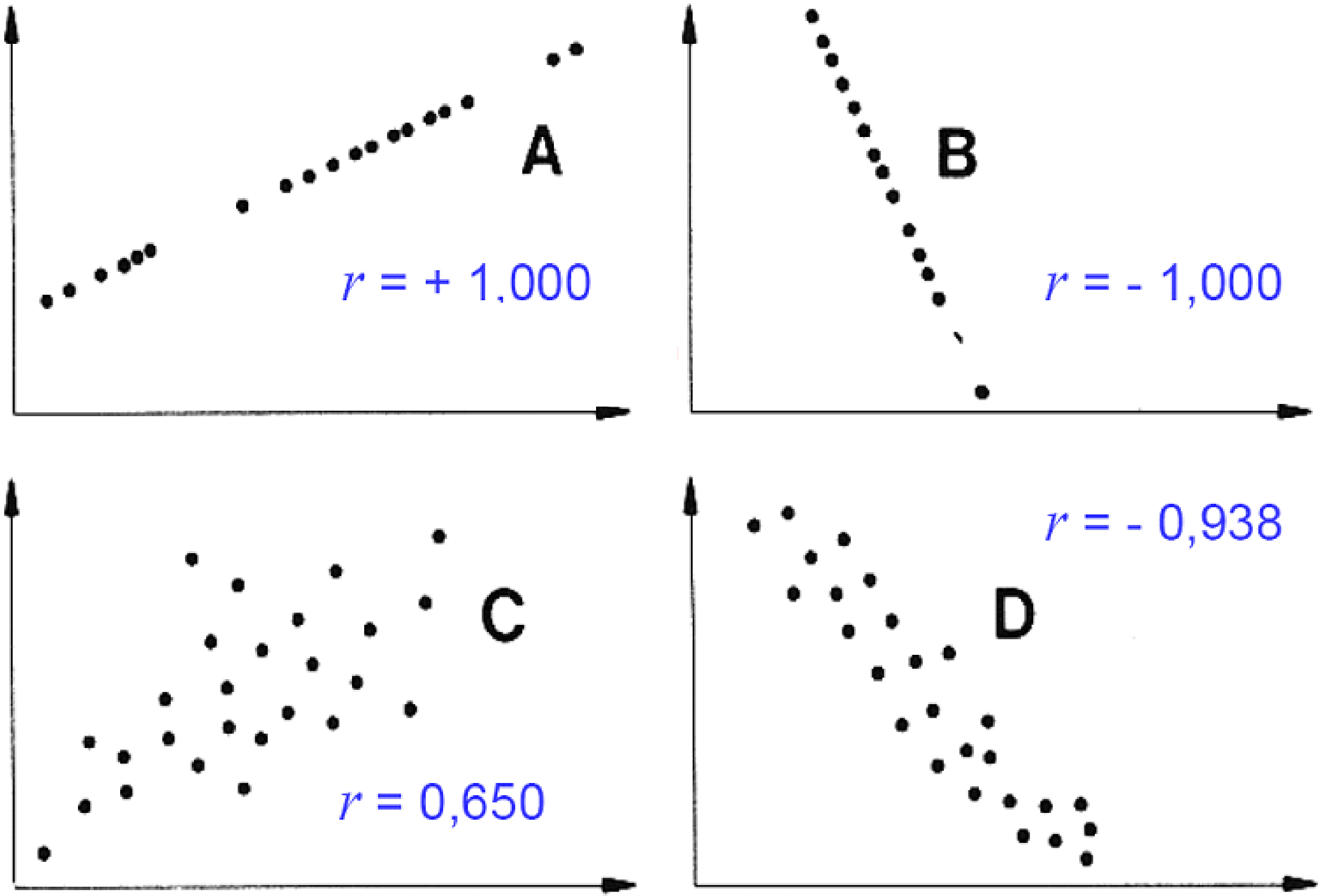


This negative association does probably not imply that a woman tends to die earlier if she gets more children.

Rather it is due to the fact, that in countries with high child mortality (and therefore low life expectancy) women tend to get more children.

An association on aggregated data does in general not imply (the same) association on the individual level!

## Examples



-> in-class exercise 1

# Pearson correlation coefficient

**If there is an exact linear relation between x and y (y=a*x+b) – regardless of the value of the steepness of the slope a - than:**

**$r_{xy}$=+/-1**

**proof:**

$$y = a\cdot x + b$$

$$\Rightarrow \overline{y} = a\cdot\overline{x} + b$$

$$y - \overline{y} = (a\cdot x + b) - (a\cdot\overline{x} + b)$$

$$\Rightarrow y - \overline{y} = a\cdot(x - \overline{x})$$

$$r_{XY} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum (X_i - \overline{X})^2}\cdot\sqrt{\sum (Y_i - \overline{Y})^2}}$$

$$= \frac{\sum (X_i - \overline{X})\cdot a\cdot(X_i - \overline{X})}{\sqrt{\sum (X_i - \overline{X})^2}\cdot\sqrt{\sum a^2 (X_i - \overline{X})^2}}$$

$$= \frac{a\cdot\sum (X_i - \overline{X})^2}{\sqrt{a^2}\cdot\sum (X_i - \overline{X})^2} = \frac{a}{|a|} = \begin{matrix} +1, & if\ a > 0 \\ -1, & if\ a < 0 \end{matrix}$$

**Always valid for a linear tranformation:**

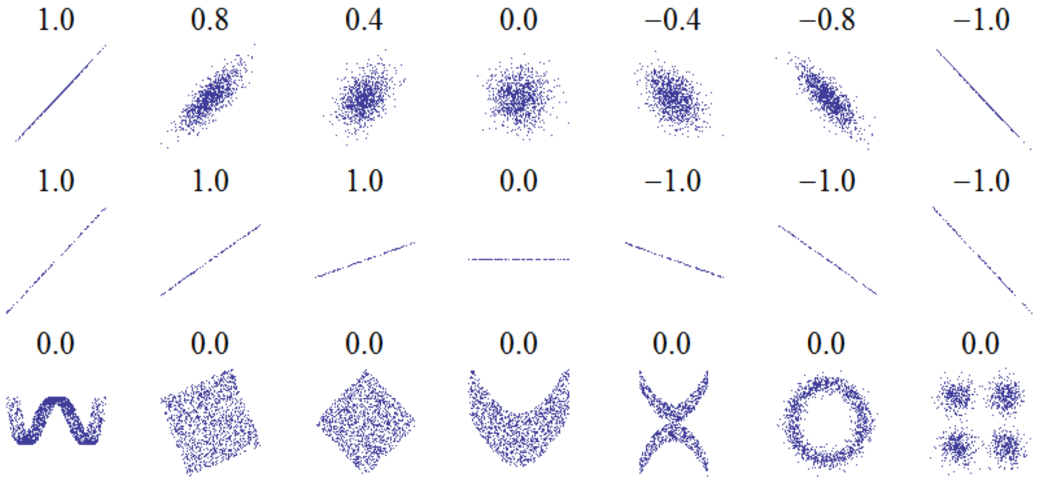
$$\overline{y} = a\cdot\overline{x} + b$$

$$sd_Y = |a|\cdot sd_X$$

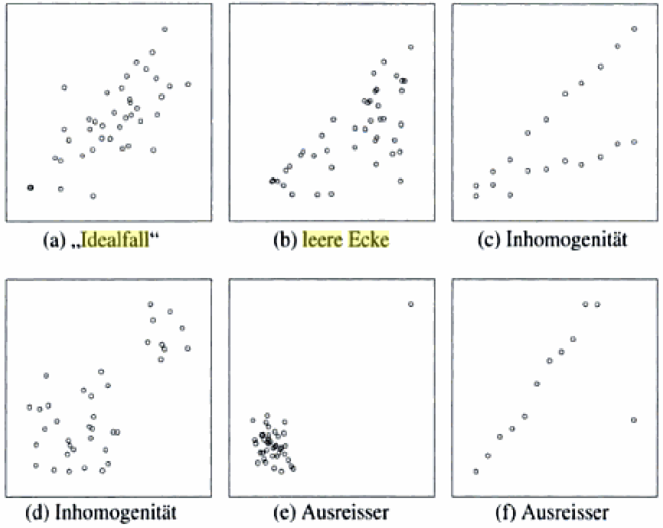
(This slide is only for proof loving people and not relevant for the exam)

# The pearson correlation is only valid for linear associations

What happens if we just calculate the correlation?



Dangerous situations



(a) „Idealfall“

(b) leere Ecke

(c) Inhomogenität

(d) Inhomogenität

(e) Ausreisser

(f) Ausreisser

Never calculate the Pearson-Correlation without inspecting the association with the help of a scatterplot.

A correlation zero does not imply that there is no association!

# Properties of the Pearson correlation

a) $-1 <= r <=1$

b) If all data points are aligned along a line with pos slope: $r=1$

   If all data points are aligned along a line with pos slope : $r=-1$

c) If there is only small scatter around a line with slope$\neq$0: r close to +/- 1

d) $r=0$, if there is no linear relation

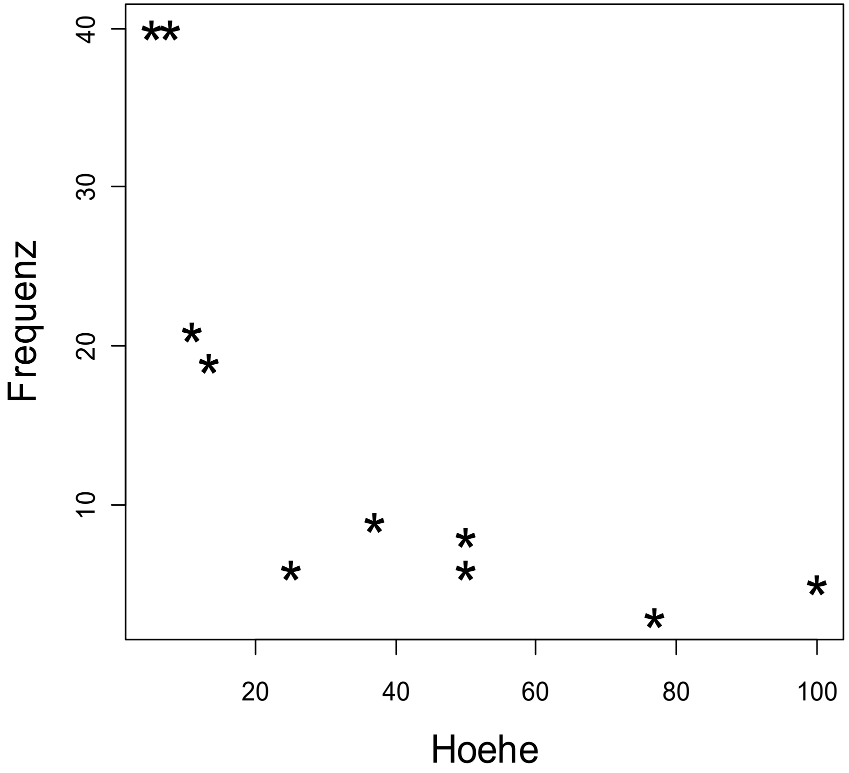
e) If $r=0$ it is still possible that there is a non-linear relation!

f ) If r is +/-1 we still can not be sure that there is a linear relation.

Always visualize your data before interpreting a correlation value!

# Spearman-Correlation
## A measure for the strength of a monotonic association

In a study (Science, 164 (1969), p.1513) the association between height of a waterfall and the frequency of the strongest ground vibration was investigated.

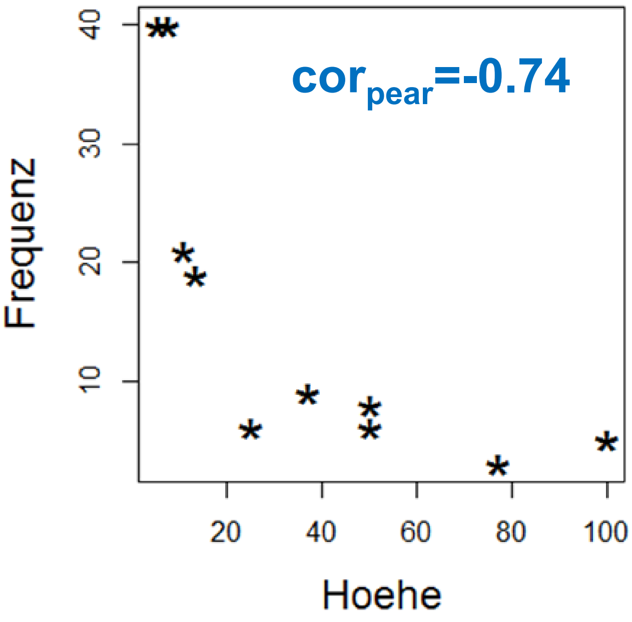


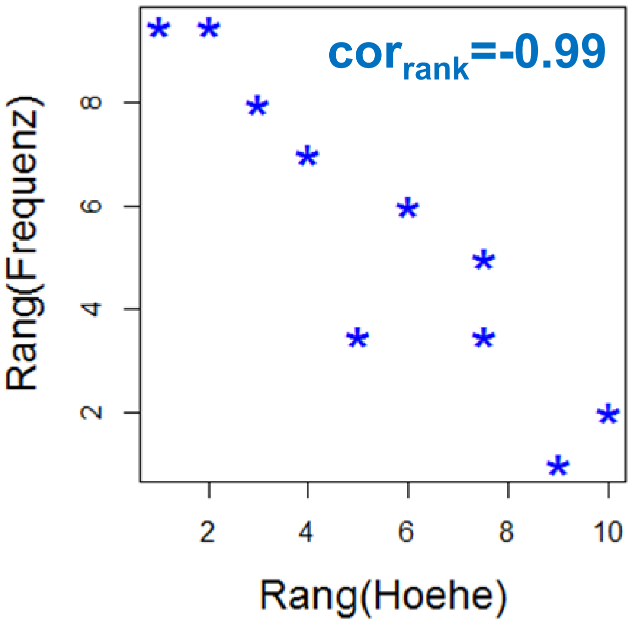
| Name | h: Hoehe | f: Frequenz | Rang(f) | Rang(h) |
|---|---|---|---|---|
| Lower.Yellowstone | 100 | 5 | | |
| Yosemite | 77 | 3 | | |
| Canadian.Niagara | 50 | 6 | | |
| American.Niagara | 50 | 8 | | |
| Upper.Yellowstone | 37 | 9 | | |
| Lower.Gullfoss | 25 | 6 | | |
| Firehole | 13.3 | 19 | | |
| Godafoss | 10.9 | 21 | | |
| Upper.Gullfoss | 7.7 | 40 | | |
| Fort.Greeley | 5.2 | 40 | | |

# Spearman-Correlation = Rank correlation

$$r_{{}^xR\,{}^yR} = \frac{\sum({}^xR_i - {}^x\overline{R})({}^yR_i - {}^y\overline{R})}{\sqrt{\sum({}^xR_i - {}^x\overline{R})^2} \cdot \sqrt{\sum({}^yR_i - {}^y\overline{R})^2}}$$

| Rang(f) | Rang(h) |
|---:|---:|
| 2 | 10 |
| 1 | 9 |
| 3.5 | 7.5 |
| 5 | 7.5 |
| 6 | 6 |
| 3.5 | 5 |
| 7 | 4 |
| 8 | 3 |
| 9.5 | 2 |
| 9.5 | 1 |





# When should we use the Spearman rank correlation?

- if there is no linear but a monotone relationship.
- if there are outliers or extreme values
- if the values ($X_i$,$Y_i$) are not bivariate Normal distributed

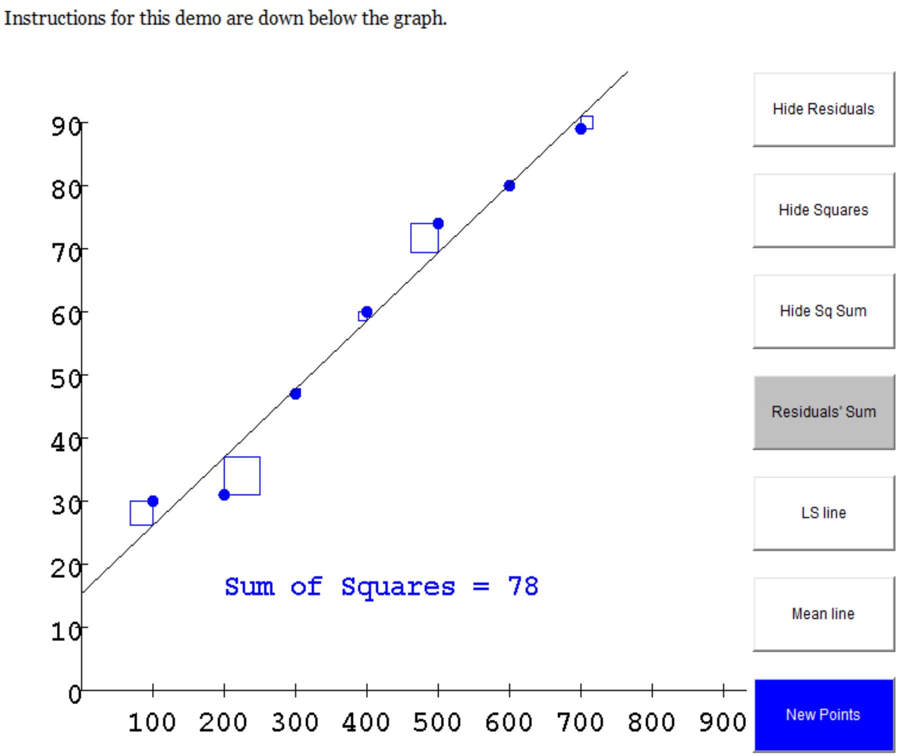
The Spearman-Correlation equals to the Pearson-Correlation applied on the ranks. Therefore, the Spearman-Correlation is robust aganinst outliers.

$$r_{{}^xR\,{}^yR} = \frac{\sum({}^xR_i - {}^x\overline{R})({}^yR_i - {}^y\overline{R})}{\sqrt{\sum({}^xR_i - {}^x\overline{R})^2} \cdot \sqrt{\sum({}^yR_i - {}^y\overline{R})^2}}$$

# Least Squares Fitting

We minimize the sum of squared residuals

$$\sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2 = \min!$$



Instructions for this demo are down below the graph.

Sum of Squares = 78

We need to fit a straight line that fits the data well.

Many possible solutions exist, some are good, some are worse.

Our paradigm is to fit the line such that the squared errors are minimized.

http://hspm.sph.sc.edu/courses/J716/demos/LeastSquares/LeastSquaresDemo.html

https://gallery.shinyapps.io/simple_regression/

Remark: According to the Gauss-Markov-Theorem the OLS (ordinary least square) fitting procedure leads to the best linear unbiased estimators (BLUE) of the regression parameters.

# Find parameter via minimizing the sum of squared residuals

$$Q(\alpha,\beta)=\sum_{i=1}^{n}(y_i-(\alpha+\beta x_i))^2=\min!$$

$$\frac{\partial Q}{\partial \alpha}=0$$

$$=\sum_{i=1}^{n}(2)(y_i-\hat{\alpha}-\hat{\beta}x_i)(-1)=0$$

$$\Rightarrow \sum_{i=1}^{n}(y_i-\hat{\alpha}-\hat{\beta}x_i)=0$$

$$\frac{\partial Q}{\partial \beta}=0$$

$$=\sum_{i=1}^{n}(2)(y_i-\hat{\alpha}-\hat{\beta}x_i)(-x_i)=0$$

$$\Rightarrow \sum_{i=1}^{n}x_i(y_i-\hat{\alpha}-\hat{\beta}x_i)=0$$

Solution:

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

$$\hat{\alpha}=\overline{y}-\hat{\beta}\overline{x}$$

# Interpretation of parameter formula

$$Y_i = a + b \cdot X_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0.\sigma^2)$$

$$\hat{y}_i = \hat{a} + \hat{b} \cdot x_i$$

slope:

$$\hat{b} = \frac{\sum_{i=1}^{n} (y_i - \overline{y}) \cdot (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)} = \frac{\operatorname{cov}(x,y)}{sd(x) \cdot sd(x)} \cdot \frac{sd(y)}{sd(y)} = \operatorname{cor}(x,y) \cdot \frac{sd(y)}{sd(x)}$$

Slope given by Δy/Δx=sd(y)/sd(x) which gets shrinked in case of non deterministic relationships

Intercept

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

Each regression line goes through center of mass

# Lineare Regression: Recap

$$Y_i = a + b \cdot X_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0.\sigma^2)$$

$$\hat{y}_i = \hat{a} + \hat{b} \cdot x_i$$

$y(\bar{x}) = \hat{a} + \hat{b}\bar{x} = \bar{y}$

Pearson Correlation

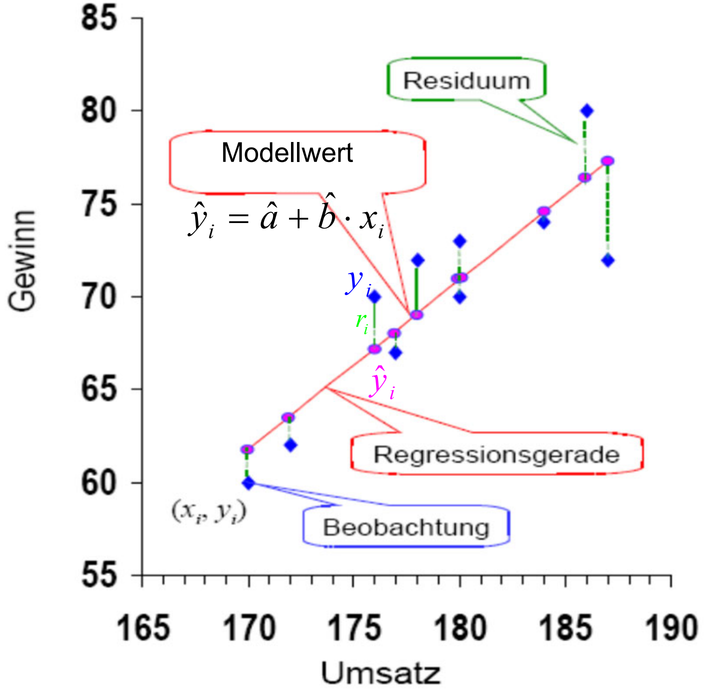
Standard-deviation of X and Y

Steigung:

$$\hat{b} = r_{xy} \cdot \frac{s_y}{s_x}$$

y-Achsenabschnitt:

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}$$



$$\hat{\sigma} = \frac{1}{n-2}\sum_{i=1}^{n} r_i^2$$

d

$$= \frac{1}{n-2}\sum_{i=1}^{n} (y_i - (\hat{a} + \hat{b} \cdot x_i))^2$$

# Correlation versus simple Regression

Both, correlation and regression investigate the association between 2 continuous variables.

For the correlation X and Y play equal roles.

For correlation we assume a bivariate Normal Distribution.

For the regression, there is no distribution assumption for X or Y but the conditional distribution (Y|X) must follow a Normal distribution.

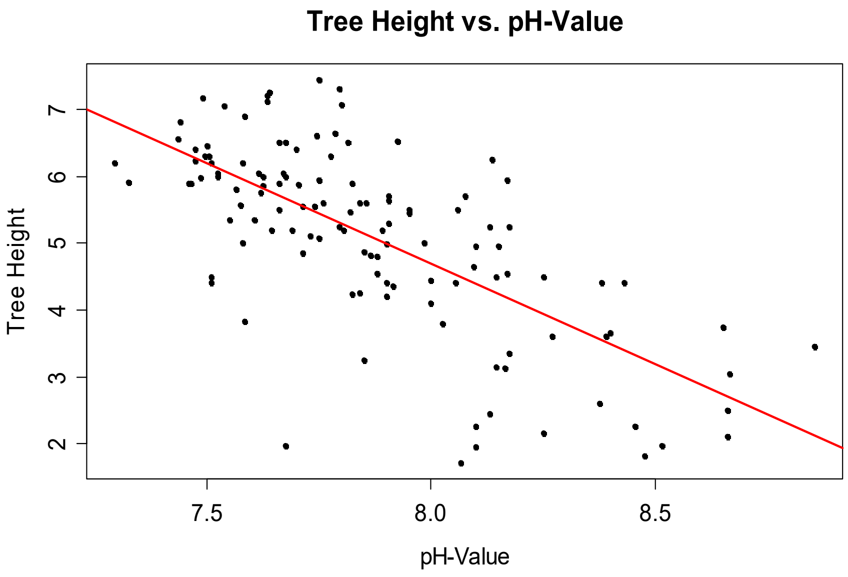
In case of regression x is the predictor variable (independent variable, explanatory variable), which is known precisely and has no error term

Y is the target variable (dependent variable, response variable, output variable), which depends on the value of X and also has an random error term imposed and therefore is a random variable.

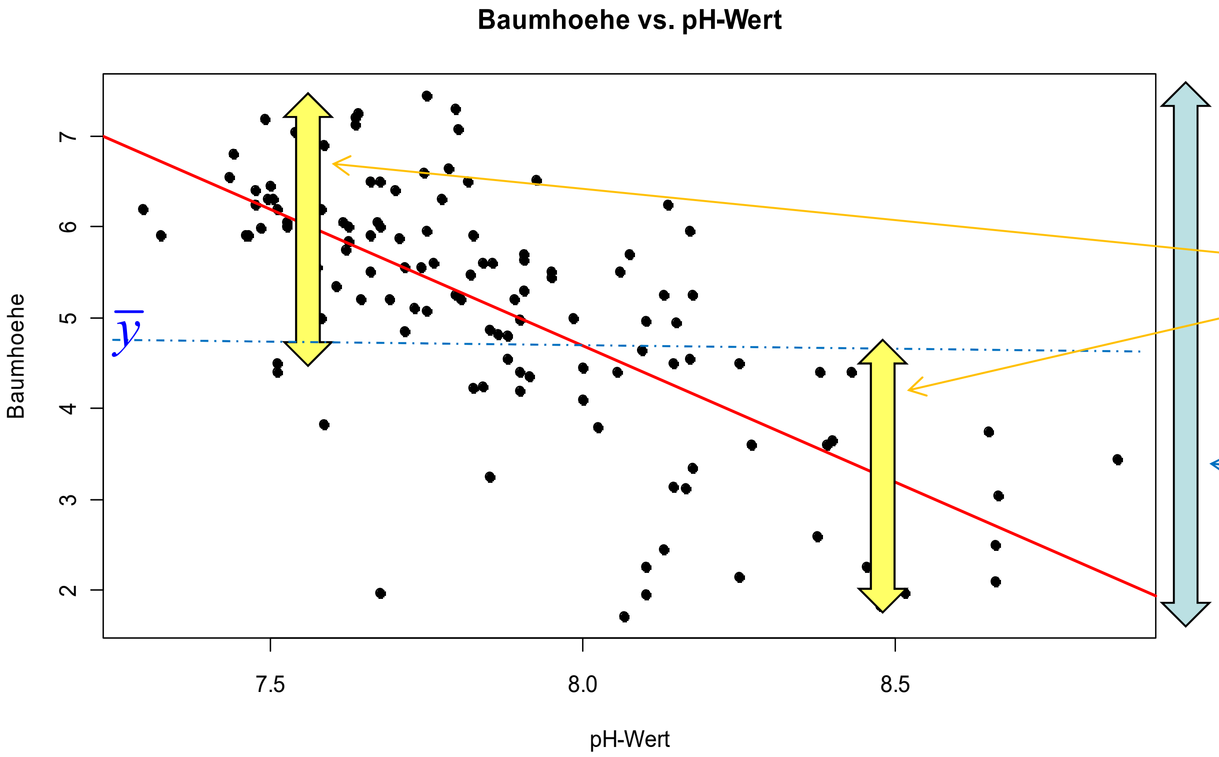
# Investigate ph effect on height of trees

```
> summary(fit)
Call: lm(formula = height ~ phvalue, data =
treeheight)

Coefficients: Estimate Std. Error t-value  Pr(>|t|)
(Intercept)   28.7227  2.2395     12.82    <2e-16 ***
phvalue       -3.0034  0.2844    -10.56    <2e-16 ***


Residual stand. err.: 1.008 on 121 degrees of freedom
Multiple R-squared: 0.4797,
```

what does it mean?



# $R^2$: How good explains the model the data?



$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}$$

We compare the sum of squared residuals to the mean with the sum of squared residuals to the fitted line.

Intuitively: the smaller the yellow range is compared to the blue one, the more useful the model is -> $R^2$ closte to 1 is good.

# $R^2$: Coefficient of Determination

If the model assumptions are fulfilled, *the R² ,named Coefficient of Determination* is often used as performance meassure. :

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} \in [0,1]$$

What is a good value for $R^2$ ? In observational studies, a value of 0.6 can mostly be considered as good. There are no formal criteria for judging this, however.

**Warning: Outliers can have high impact on $R^2$ :**

**always perform a residual analysis.**

# Investigate pollution effect on mortality by regression

In an observational study mortality rates and many possible predictors were collected. Here we only use three of the predictors – we are interested in the effect of SO2 and adjust for the influence of the other two predictors.

```
fit = lm( Mortality ~ log(SO2) + NonWhite + Rain, data=mortality)

summary(fit)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 773.0197    22.1852  34.844  < 2e-16 ***
log(SO2)     17.5019     3.5255   4.964 7.03e-06 ***
NonWhite      3.6493     0.5910   6.175 8.38e-08 ***
Rain          1.7635     0.4628   3.811 0.000352 ***
---
Residual standard error: 38.4 on 55 degrees of freedom
Multiple R-squared: 0.641,  Adjusted R-squared: 0.6214
F-statistic: 32.73 on 3 and 55 DF,  p-value: 2.834e-12
```

what does it mean?

Remark: Before making any interpretation we should check if the assumptions for the linear regression model are not violated.

# adjR$^2$: Adjusted Coefficient of Determination

- If we add more and more predictor variables to the model, R-squared will always increase, and never decreases
- **We should adjust for the number of predictors!**

$$adjR^2 = 1 - \frac{n-1}{n-(p+1)} \cdot \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \in [0,1]$$

# What does the F-value and the global p-Value in `lm` mean?

*Question*: is there ***any*** relation between predictors and response?

We test the null hypothesis (the mean alone is already a good model)

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

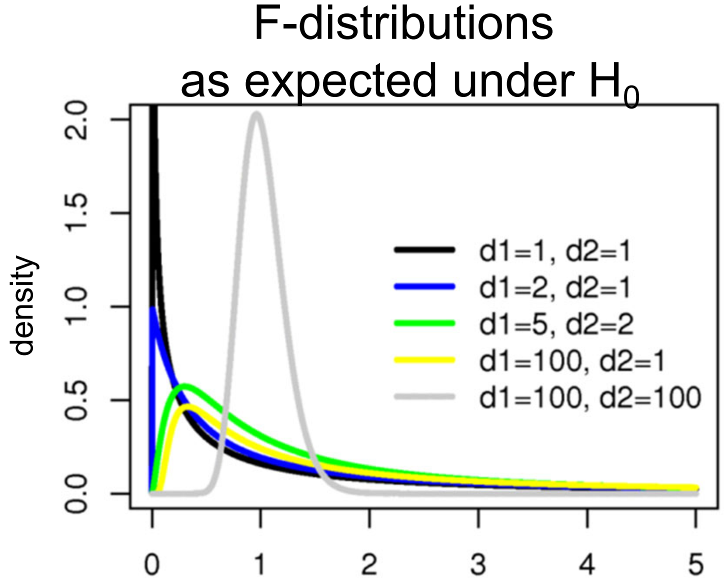
against the alternative (we need at least 1 predictor)

for at least one j in 1,…, p

$$H_A : \beta_j \neq 0$$

The test statistic is:

$$F = \frac{n-(p+1)}{p} \cdot \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \overset{unter\ H_0}{\sim} F_{p,n-(p+1)}$$


F-distributions as expected under $H_0$

If the F-Value calculated from the fits get to big (> ${}^{95\%}q_F$), then the p-value (area under density right from $F_{got}$) get small and we can reject $H_0$.

# Model selection: compare nested model?

- *Question*: does the model improve significantly if I include more predictors?
- We test the $H_0$: the smaller model with j predictors is already good

```
fit.small = lm( y ~ x1 + x2 +…+ xj, data=my.dat)
```

- against $H_A$: we need a bigger model with (k-j) additional predictors

```
fit.big = lm( y ~ x1 + x2 +…+ xj +…+ xk, data=my.dat)
```

- The test statistic to compare the performance of both models is based on the unadjusted $R^2$-values of the fitted models:

$$F = \frac{n-k}{k-j} \cdot \frac{R_k^2 - R_j^2}{1-R_k^2} \overset{unter\ H_0}{\sim} F_{k-j,n-k}$$

```
anova(fit.big, fit.small) # as result we get the F- and p-value
```

In **anova** the sum of squared residuals are compared within the different groups or models and a F-Value is determined – if  we get a big F-value and a small p-value (>5%), we can reject $H_0$ and conclude that the bigger model is significantly better.

# Variable Selection

**Goal:** We want to develop a simple model by dropping all predictors f from the regression model which are not necessary.

**How:** In a step-by-step manner, e.g. the least significant predictor is dropped from the model, as long as we have significant p-values.

**In R:**

```
> fit <- update(fit, . ~ . - colx)
> summary(fit)
```

**Warning:** The p-values of the individual hypothesis tests are based on the assumption that the other predictors remain in the model and do not change. Therefore, you must not drop more than one single non-significant predictor at a time! Moreover, after variable selection the remaining coefficients and p-values are biased leading to an overestimation of effect size and significance.

# Main pitfalls when selecting variables for a linear regression model

- Variable selection can lead to
  - biased parameter estimates
  - biased p-values
- Including collider-variables lead to distorted associations

# Why coefficients estimates are not unbiased after model selection

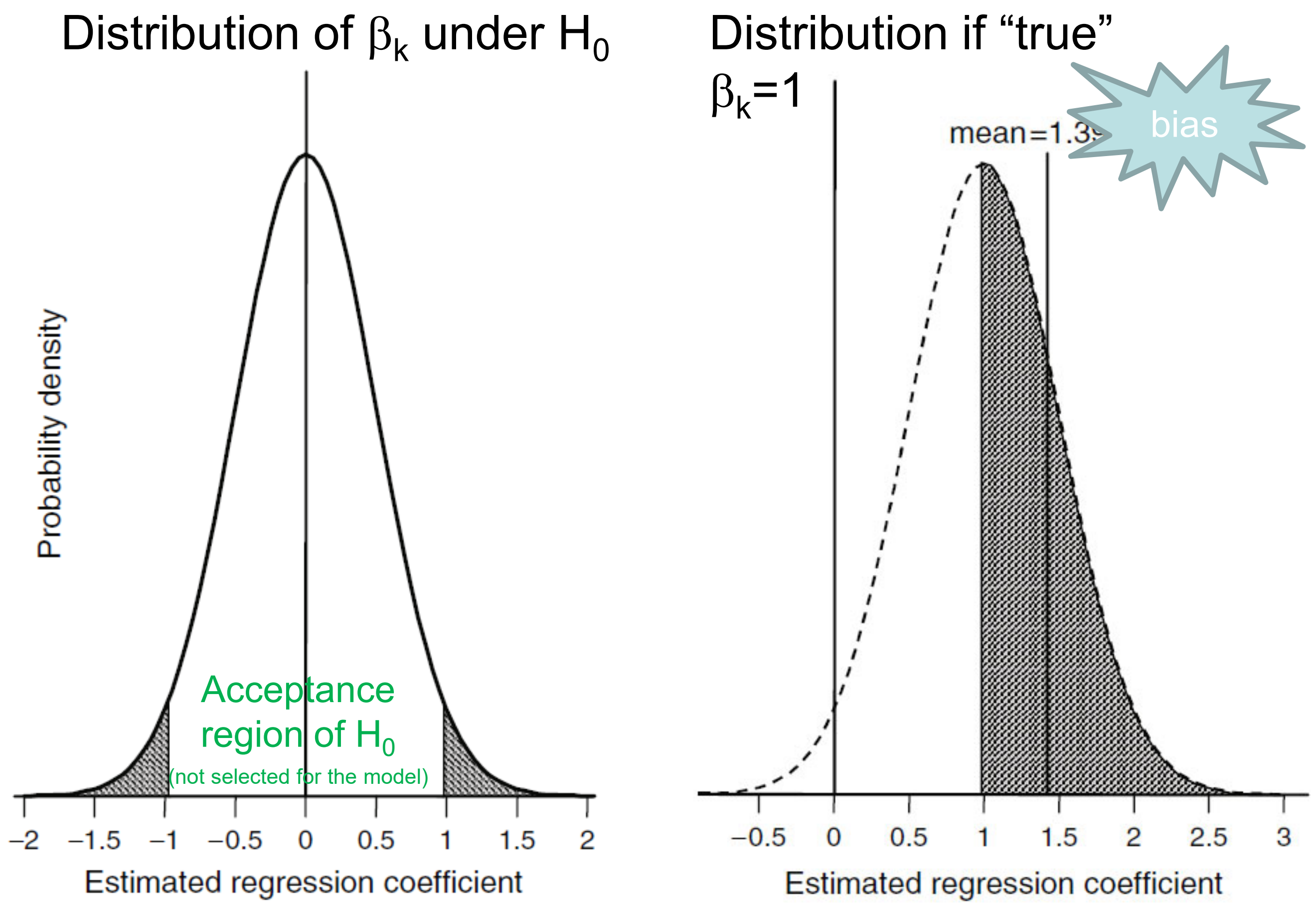


**Fig. 5.5** Illustration of testimation bias. In case of a noise variable, the average of estimated regression coefficients is zero, and 2.5% of the coefficients is below −0.98 (1.96 × SE of 0.5), and 2.5% of the coefficients is larger than +0.98 (1.96 × SE of 0.5). In case of a true coefficient of 1, the estimated coefficients are statistically significant in 52%. For these cases, the average of estimated coefficients is 1.39 instead of 1

Picture taken from E.W. Steyerberg "Clinical Prediction Models" 2009

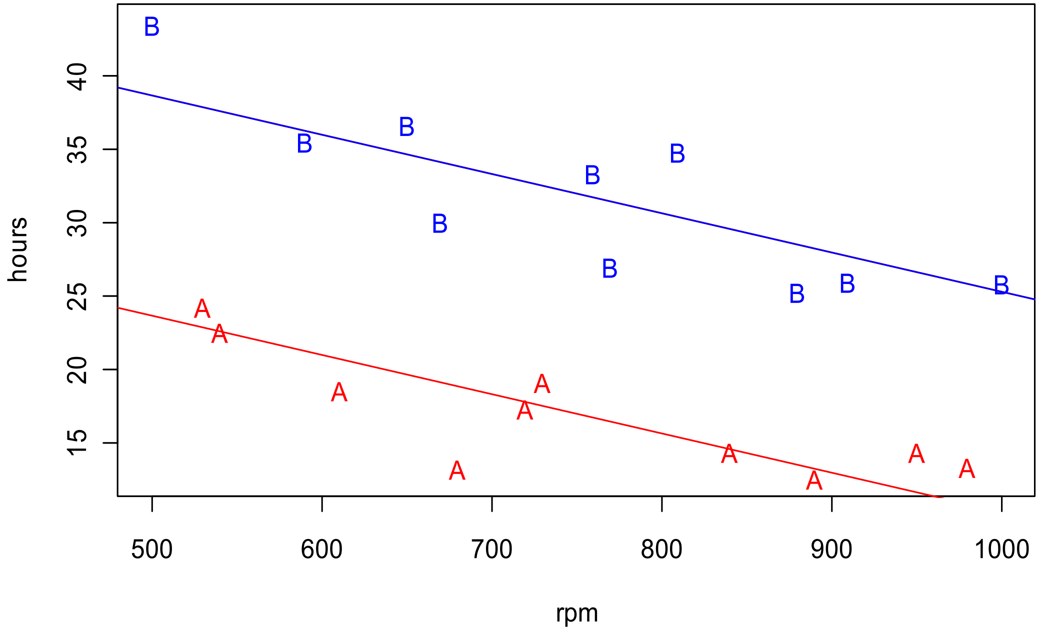
# Linear Regression with continuous and factorial predictors

**Output:** **hours:** lifetime of a cutting tool

**Predictor 1:** **rpm**: speed of the machine in rpm

**Predictor 2:** **tool:** tool type A or B

```
fit1 <- lm(hours ~ rpm + tool, data=my.dat)
```

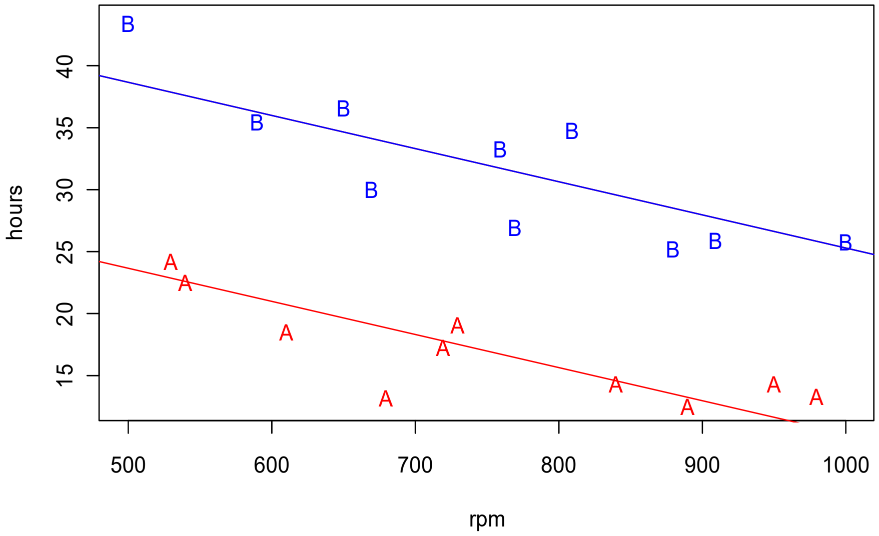


We have an additive model: the difference between the tools is a shift.

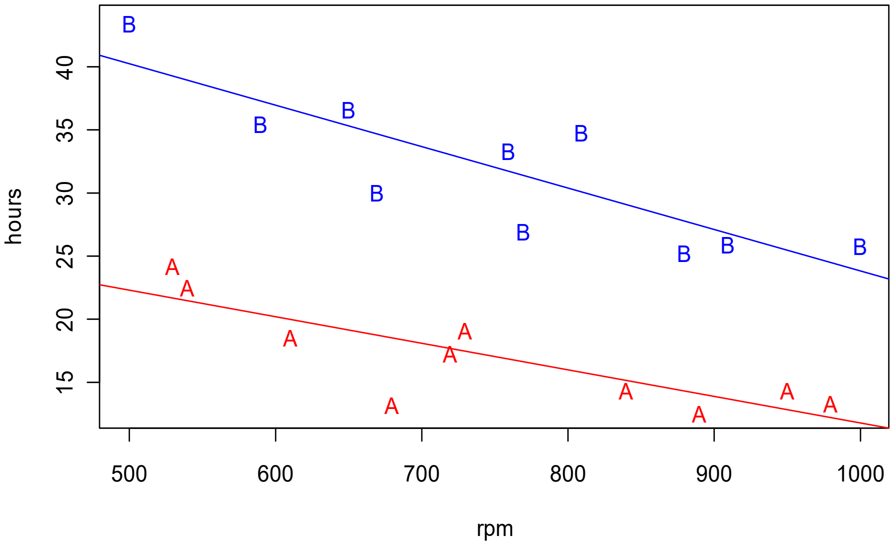
# What does interaction mean?

## Different slopes of continuous variables at different levels of a factor

Do not allow for interaction



Interaction as allowed



```
fit1=lm(hours ~ rpm + tool,
                data=my.dat)
```

```
fit2=lm(hours ~ rpm * tool,
                data=my.dat)
```

In case of interaction, the slope of the predictor “rpm” changes for different levels of the second predictor “tool”.

# Do we get the same slope in rpm for tool A and tool B? Is there an interaction between rpm and tool?

```
fit2 <- lm(hours ~ rpm * tool, data=my.dat)

> summary(fit2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.774760   4.633472   7.073 2.63e-06 ***
rpm         -0.020970   0.006074  -3.452  0.00328 **
toolB       23.970593   6.768973   3.541  0.00272 **
rpm:toolB   -0.011944   0.008842  -1.351  0.19553
---
Residual standard error: 2.968 on 16 degrees of freedom
Multiple R-squared: 0.9105,  Adjusted R-squared: 0.8937
F-statistic: 54.25 on 3 and 16 DF,  p-value: 1.319e-08
```

$$\text{hour} = 32.8 + -0.02 \cdot \text{rpm} + 24 \cdot \text{toolB} \text{ -}0.01 \cdot (\text{rpm} \cdot \text{toolB})$$

The main effects are hard to interpret in case of interactions.

Here the interactions seems not to be significant. With ANOVA we can test for nested models if the more complex model leads to a significant improvement:

# How to read a model with interaction?

$$\text{hour} = 32.8 - 0.02 \cdot \text{rpm} + 24 \cdot \text{toolB} - 0.01 \cdot (\text{rpm} \cdot \text{toolB})$$

toolB (toolB=1):

$$\text{hour} = 32.8 - 0.02 \cdot \text{rpm} + 24 \cdot 1 - 0.01 \cdot (\text{rpm} \cdot 1)$$

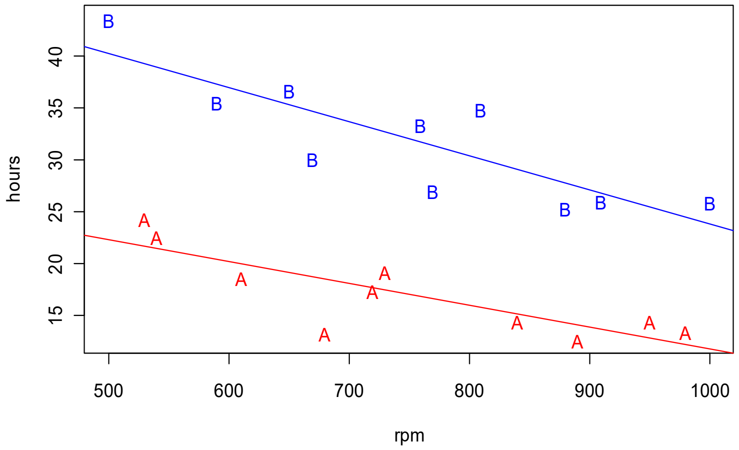
$$\text{hour} = 56.9 - 0.03 \cdot \text{rpm}$$

toolA (toolB=0):

$$\text{hour} = 32.8 - 0.02 \cdot \text{rpm} + 24 \cdot 0 - 0.01 \cdot (\text{rpm} \cdot 0)$$

$$\text{hour} = 32.8 - 0.02 \cdot \text{rpm}$$

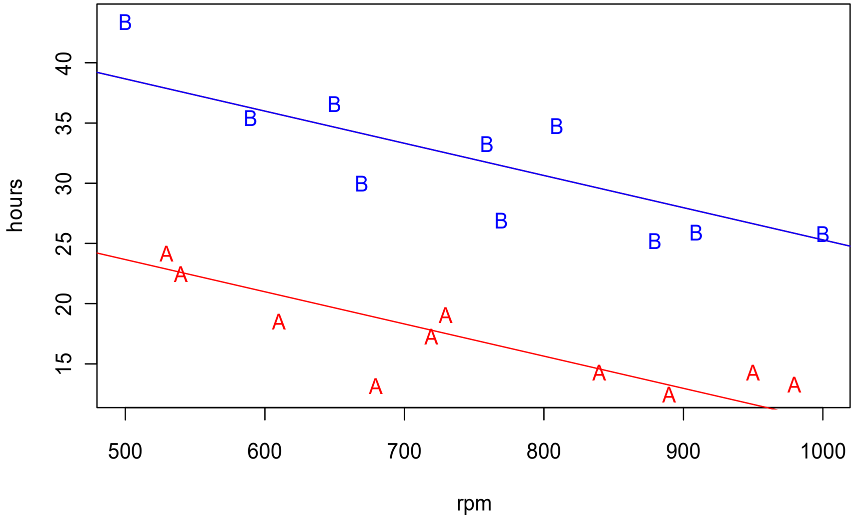
Interaction is allowed



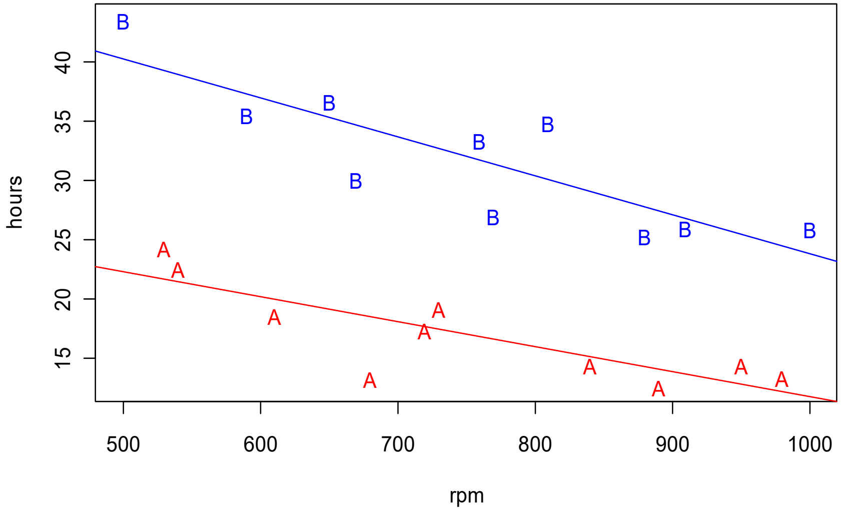
In case of interaction, the slope of the predictor “rpm” changes for different levels of the second predictor “tool” – also the intercept is changing for the two tools.

Remark: In case of interaction between two continuous predictors, slope (and intercept) of one predictor changes continuously with a continuous changing value of the other predictor and vice versa.

# Do we need the complex model with the interaction?



Do not allow for interaction



Interaction is allowed

```
fit1=lm(hours ~ rpm + tool,
             data=my.dat)
```

```
fit2=lm(hours ~ rpm * tool,
             data=my.dat)
```

```
anova(fit2,fit1)
# p>5%, therefore interaction is not needed
```