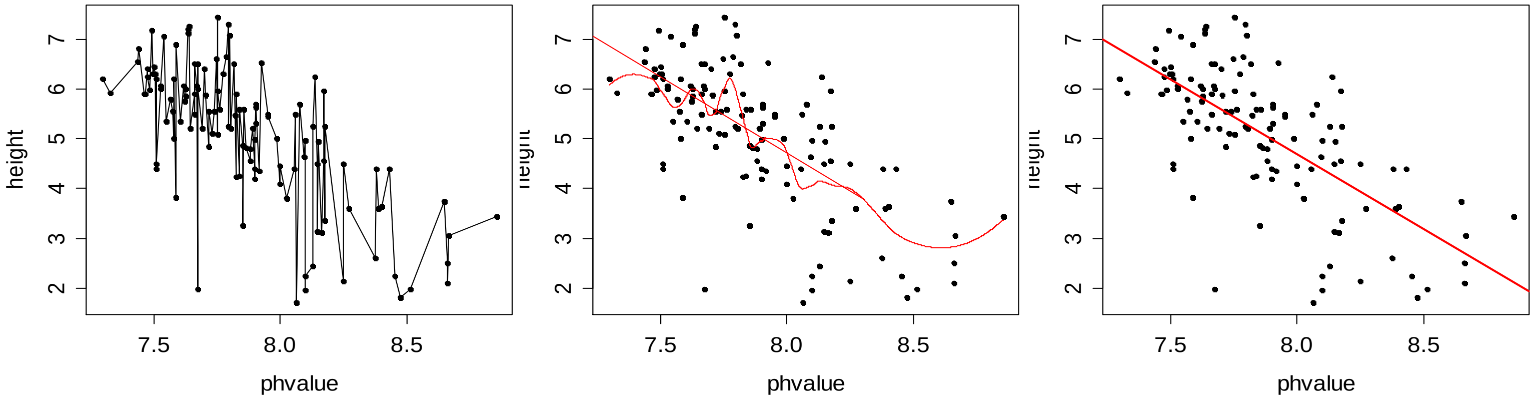
Biostatistics
Week 8

- **Correlation between two continuous measures**
  - Pearson correlation for linear relations
  - Spearman rank correlation for monotone relations

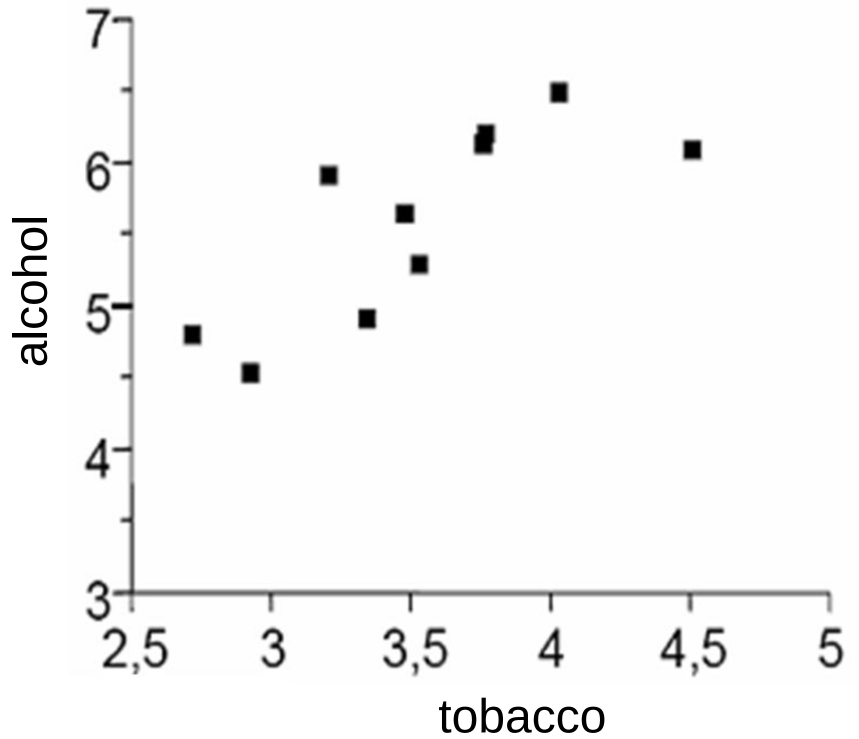
- **Simple Ordinary Least Square Regression**
  - model assumptions
  - model fitting, parameter estimation
  - interpretation of a regression model



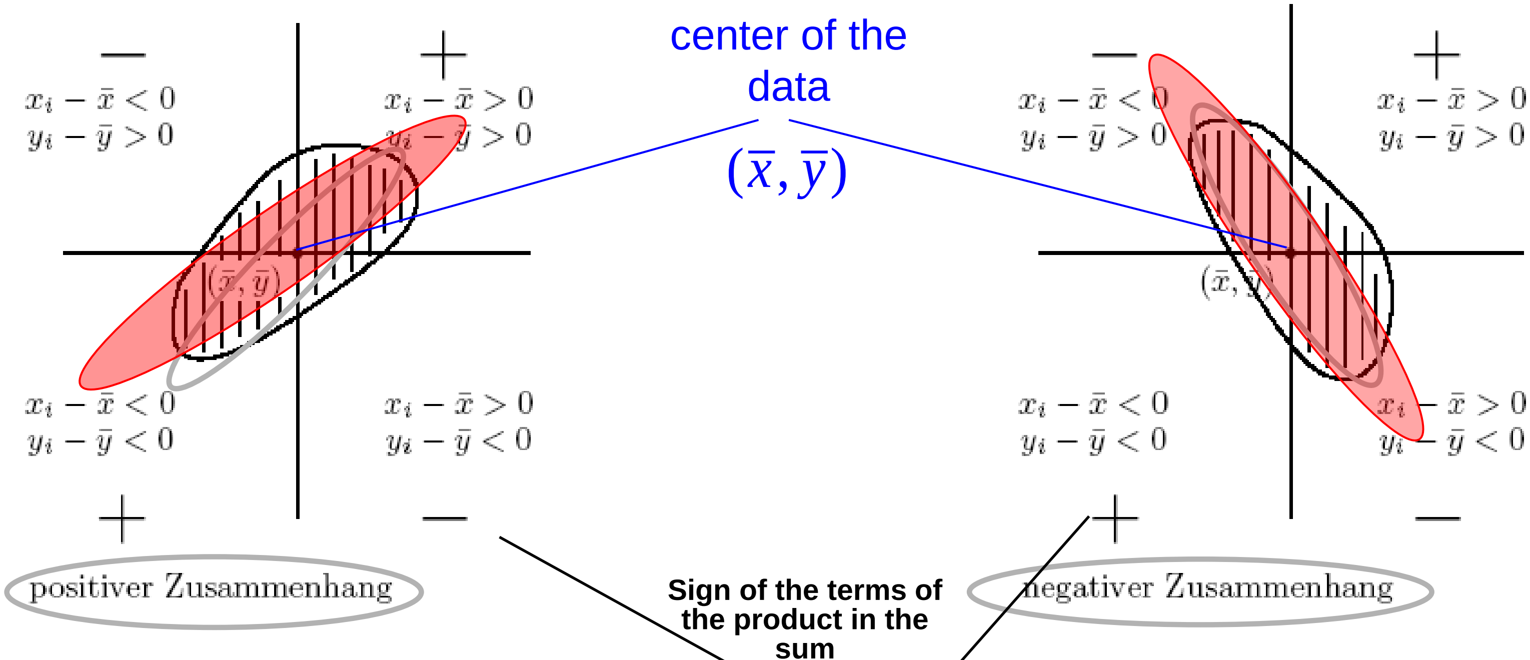
Is there an association between 2 variables?

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



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).

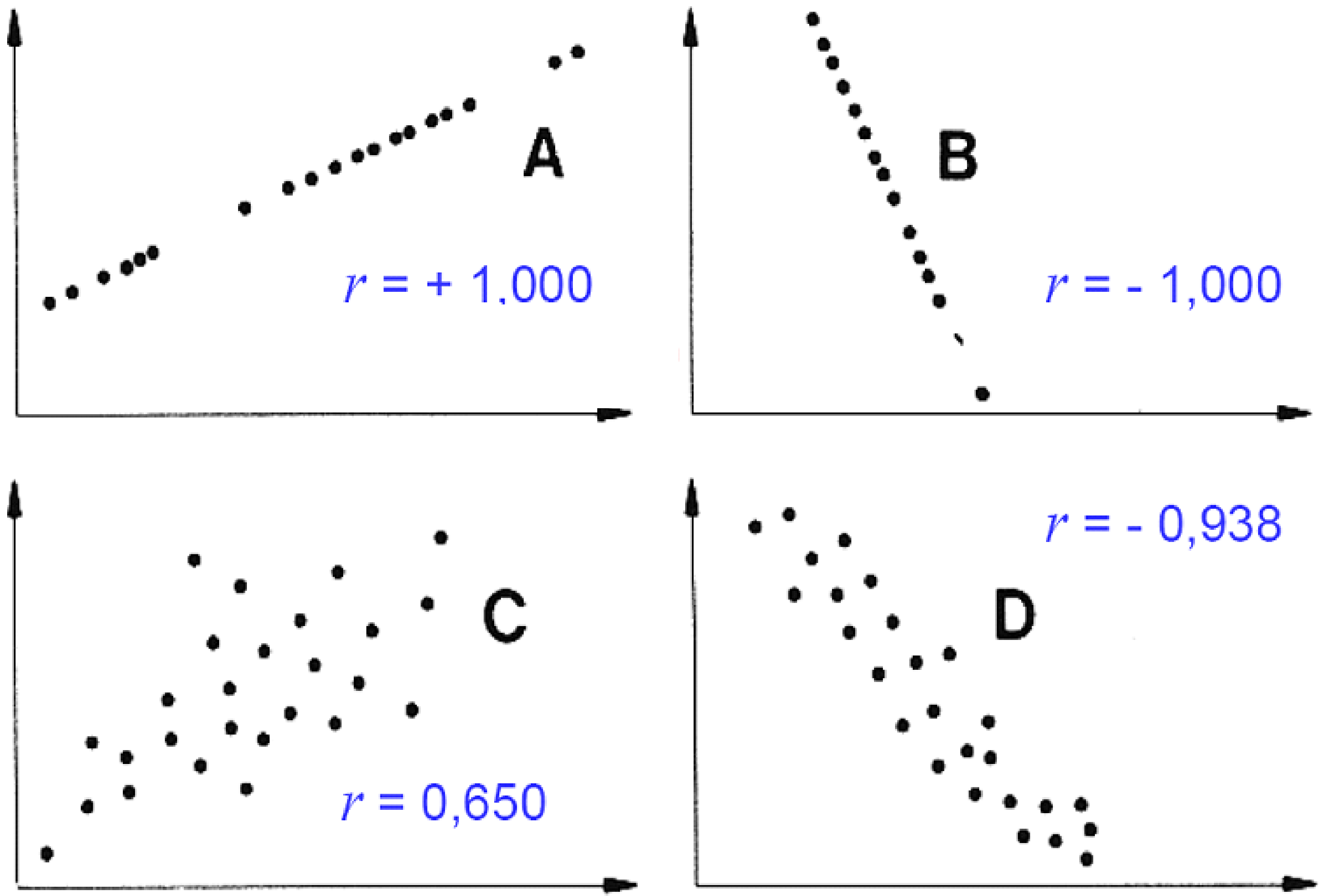
In a scatterplot y vs x or x vs y that means how closly the points scatter around a imagined straight line.

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)

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:

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

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

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

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

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

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

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

Always valid for a linear tranformation:

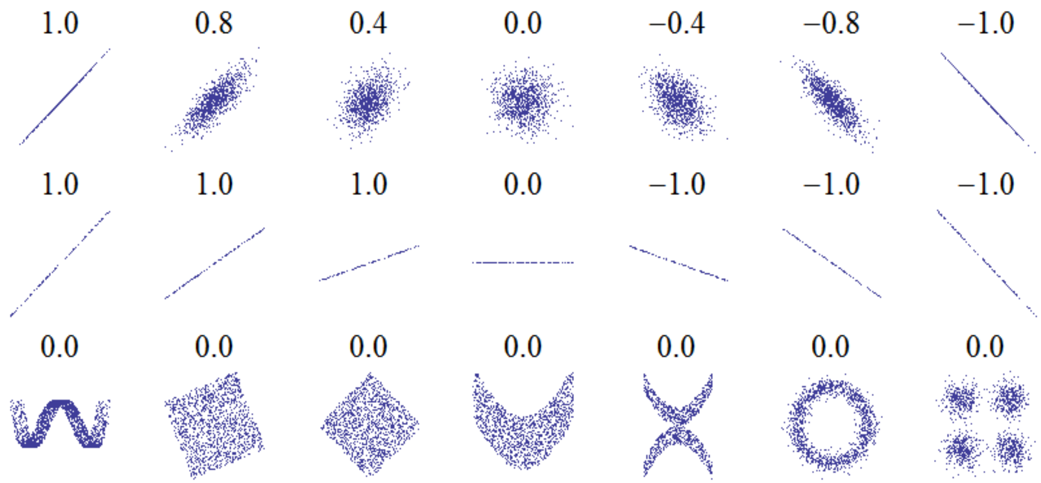
$$\bar{y} = a \cdot \bar{x} + b$$

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

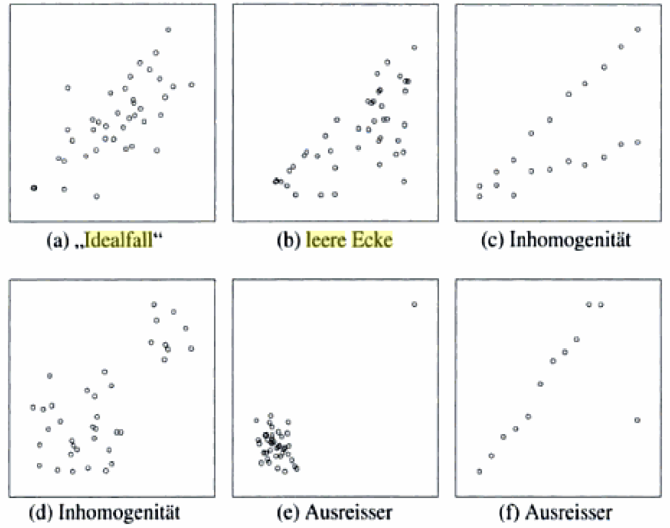
(This slide is only for proof loving people and not relevant for the exam)

The pearson correlation is only valid for linear associations

What happens if we just calculate the correlation?



Dangerous situations



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≠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!

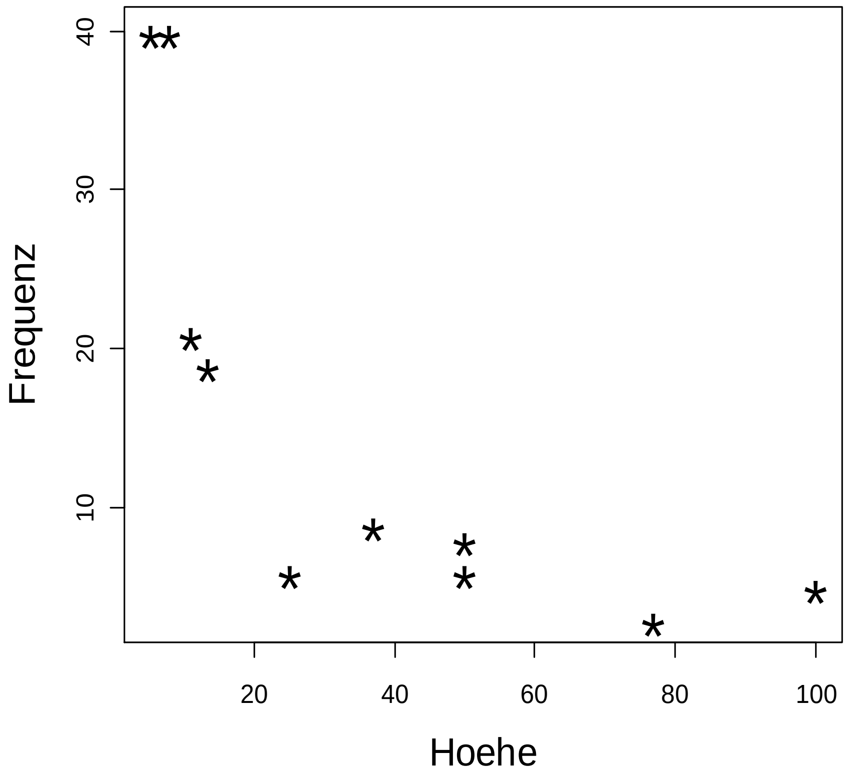
f) If r is +/-1 we still can not be sure that there is a linear relation.

Always visualize your data before interpreting a correlation value!

Spearman-Correlation

A measure for the strength of a <u>monotonic</u> 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.

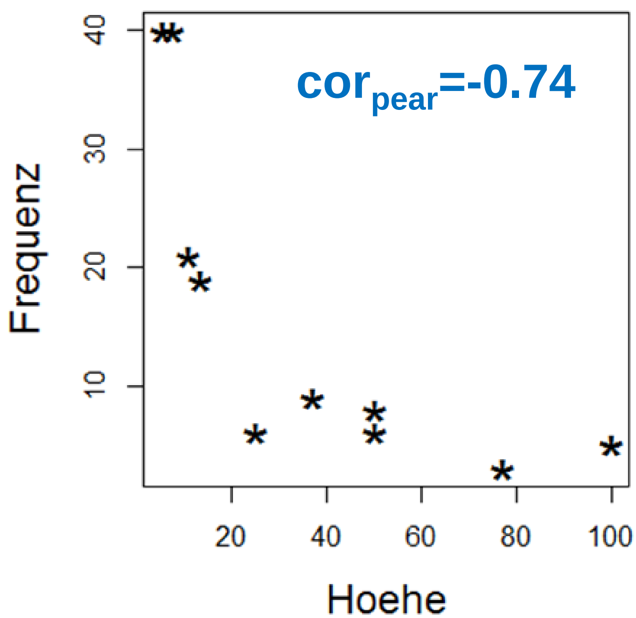


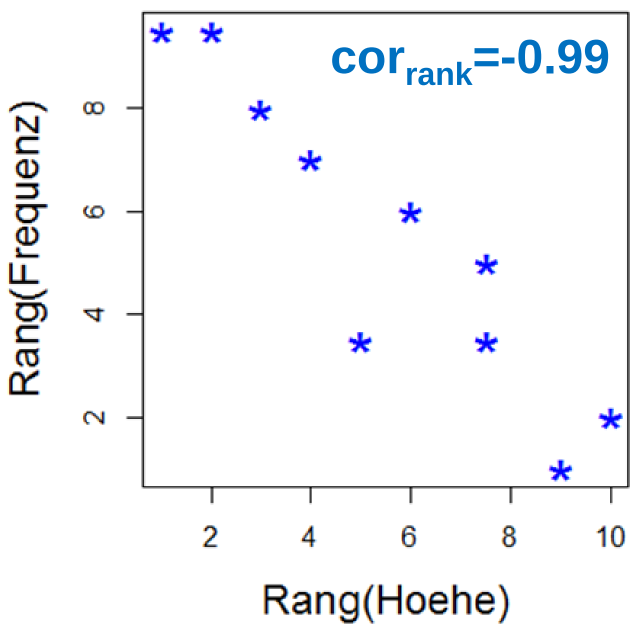
Name	h: Hoehe	f: Frequenz	Rang(f)	Rang(h)
Lower.Yellowstone	100	5		
Yosemite	77	3		
Canadian.Niagara	50	6		
American.Niagara	50	8		
Upper.Yellowstone	37	9		
Lower.Gullfoss	25	6		
Firehole	13.3	19		
Godafoss	10.9	21		
Upper.Gullfoss	7.7	40		
Fort.Greeley	5.2	40		

Spearman-Correlation = Rank correlation

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

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.

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

CAN WE PREDICT A STUDENT'S WEIGHT y FROM HIS OR HER HEIGHT x?

Regression analysis

FITS A STRAIGHT LINE TO THIS MESSY SCATTERPLOT. x IS CALLED THE ***INDEPENDENT*** OR ***PREDICTOR*** VARIABLE, AND y IS THE ***DEPENDENT*** OR ***RESPONSE*** VARIABLE. THE ***REGRESSION*** OR ***PREDICTION*** LINE HAS THE FORM

$$y = a + bx$$

The Cartoon Guide to Statistics,
Larry Gonick and Woollcott Smith

When do we use regression?

Everyday question:

How does a continuous target variable of special interest depend on several other (explanatory) factors.

Examples:

- growth of plants, affected by fertilizer, soil quality, …
- costs per patient, affected by diagnose, age, of the patient, …

Regression:

- quantitatively describes relation between predictors and target
- high importance, most widely used statistical methodology

Goals of Linear Modeling

Goal 1: Model the relations, interpretation of the parameters

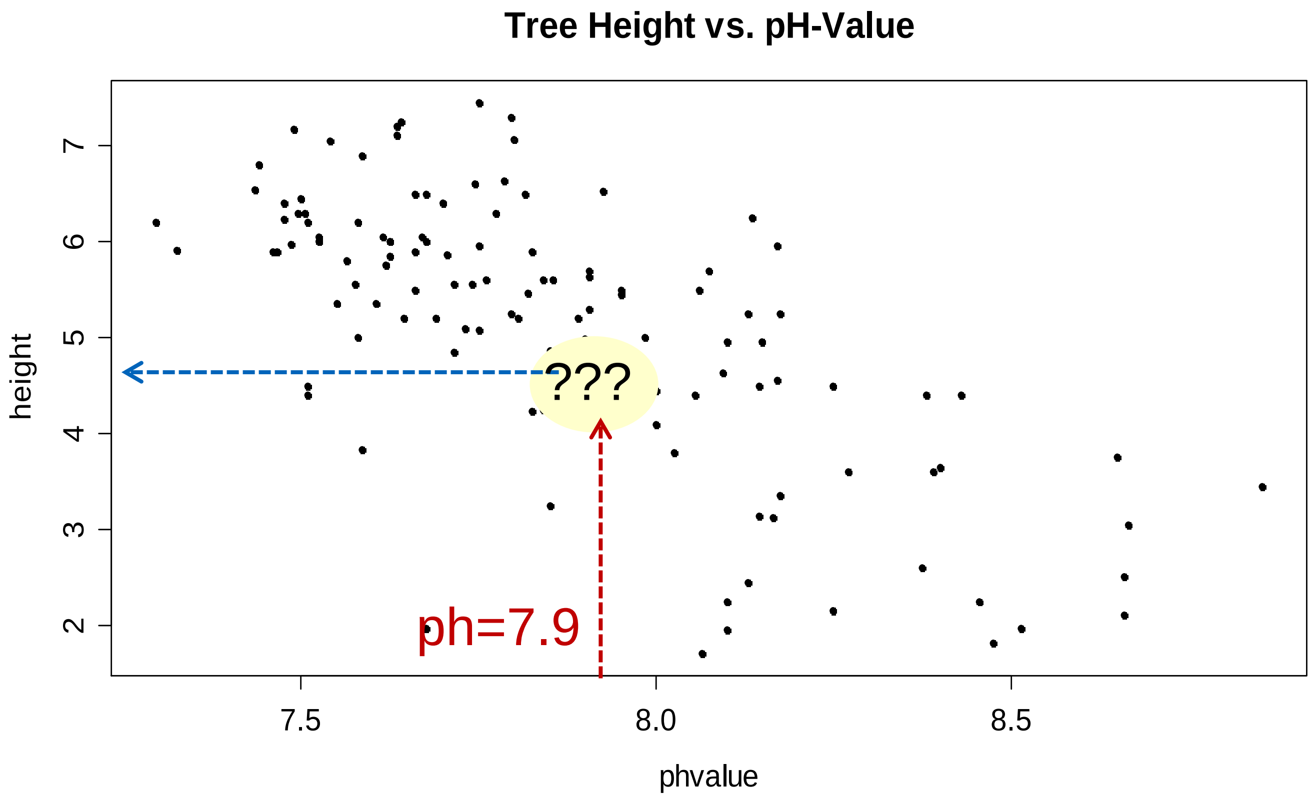
- Does a fertilizer positively associated with plant growth?
- Regression is a tool to give an answer on this
- However, showing causality is a different matter

Goal 2: Target value prediction for new explanatory variables

- Which value do we expect for the bone elasticity of a certain mouse?
- It also provides an idea on the uncertainty of the prediction

A continuous variable as explanatory variables

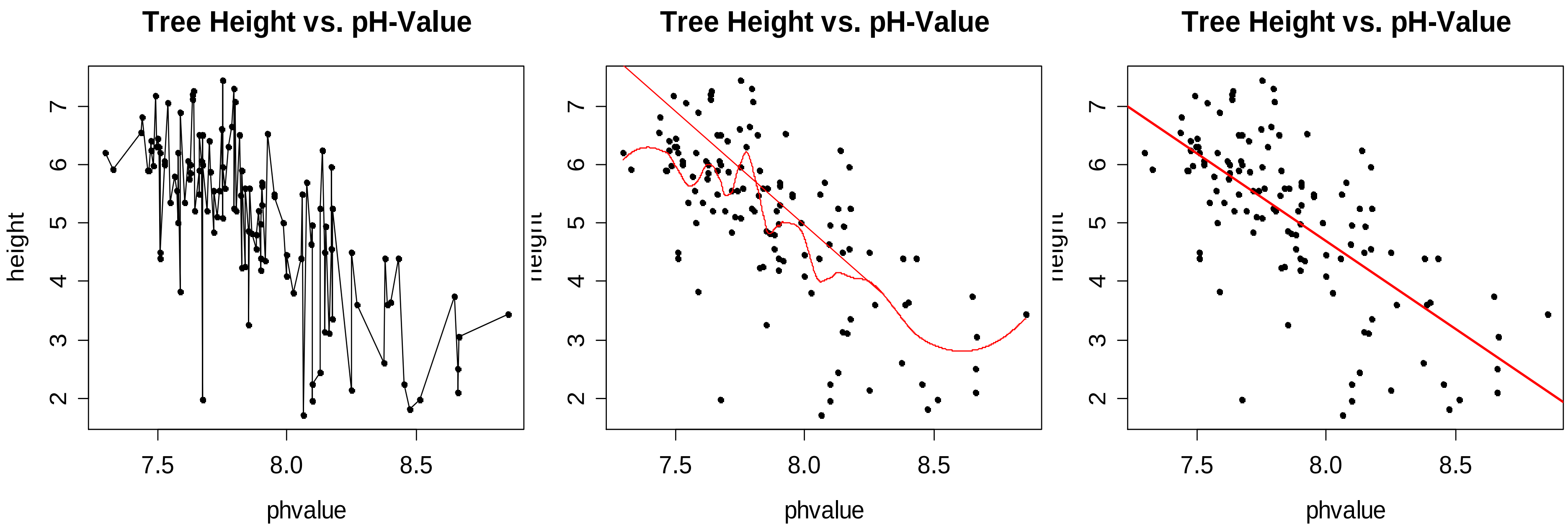
Task: How does tree Height depend on pH-value?



Which height would we expect at ph=7.9?

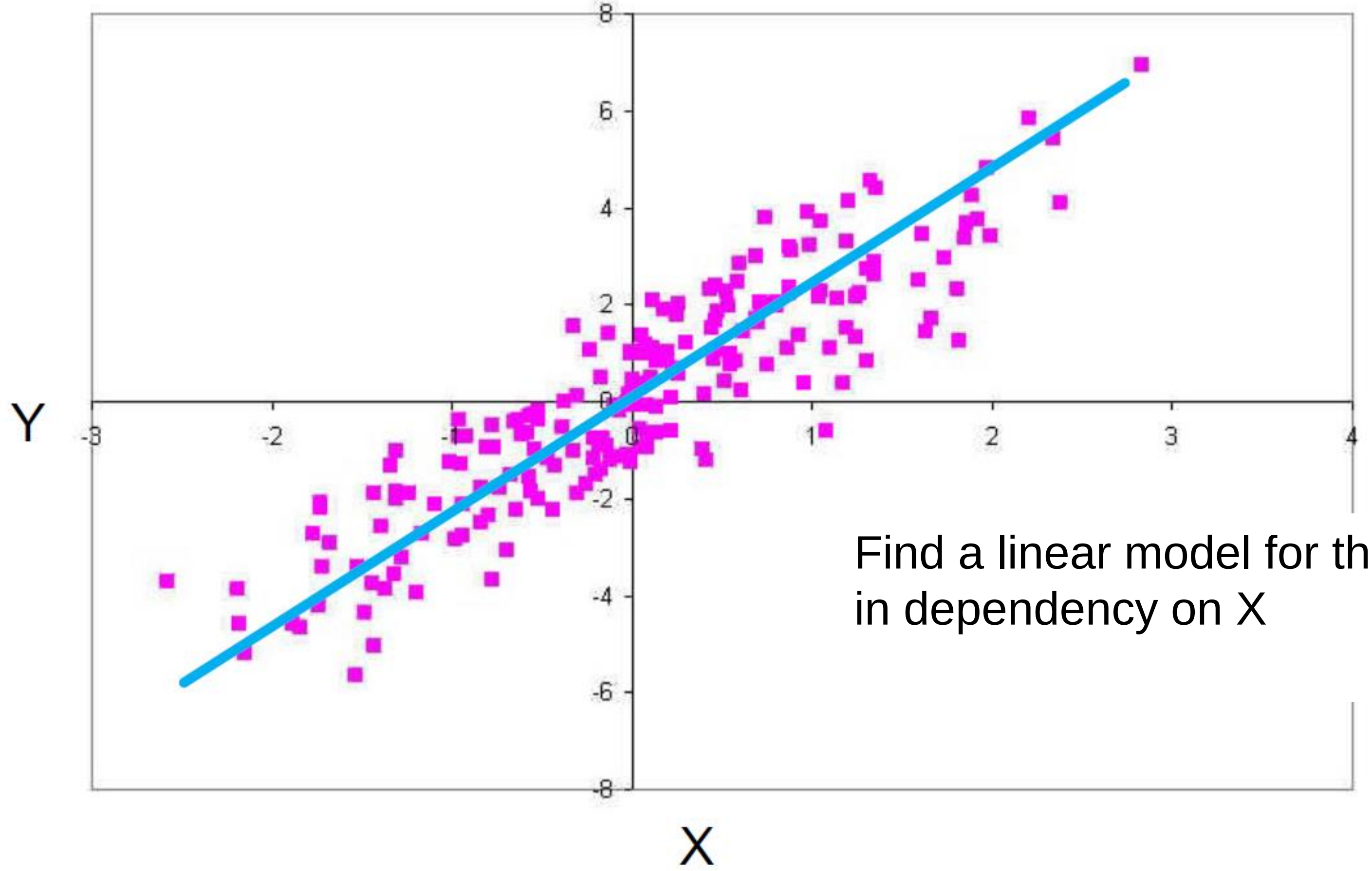
Describing the relation: What is a good model?

What is a good model for the relation between pH-value and tree height?



Remark: The first model fits the training data perfect but does probably over-fit the data. To evaluate the prediction performance of a model without using model theory we can use cross-validation: leave out successively each data point, determine the model with remaining data and use the model to predict left out value. The model is best which produces the best predictions on new or left out data points.

Simple linear regression: Only one explanatory variable



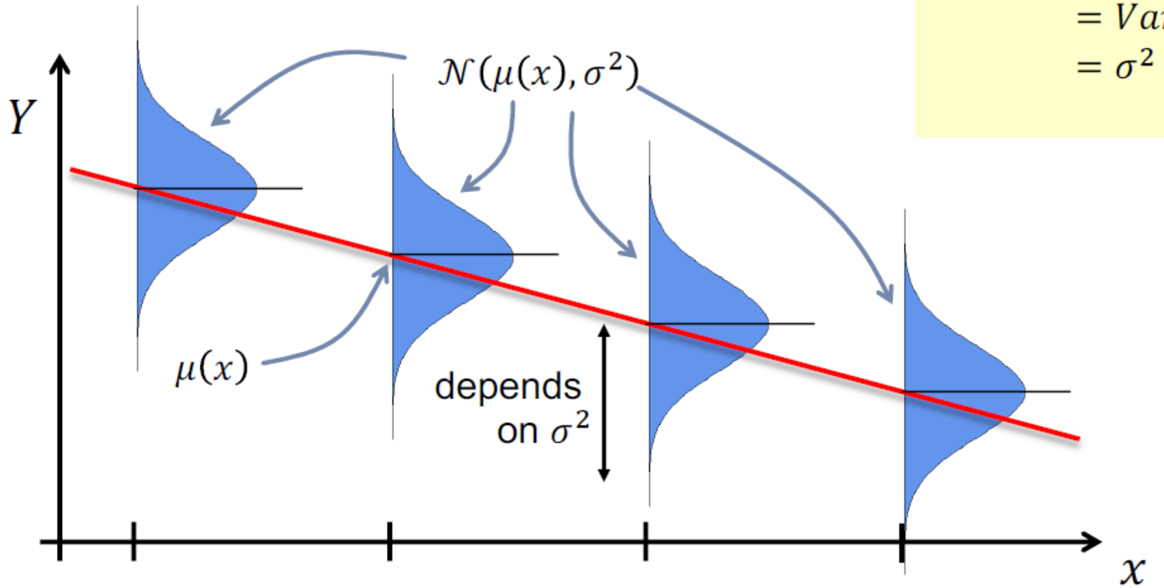
Find a linear model for the mean of Y in dependency on X

Linear regression: Two possible model definitions

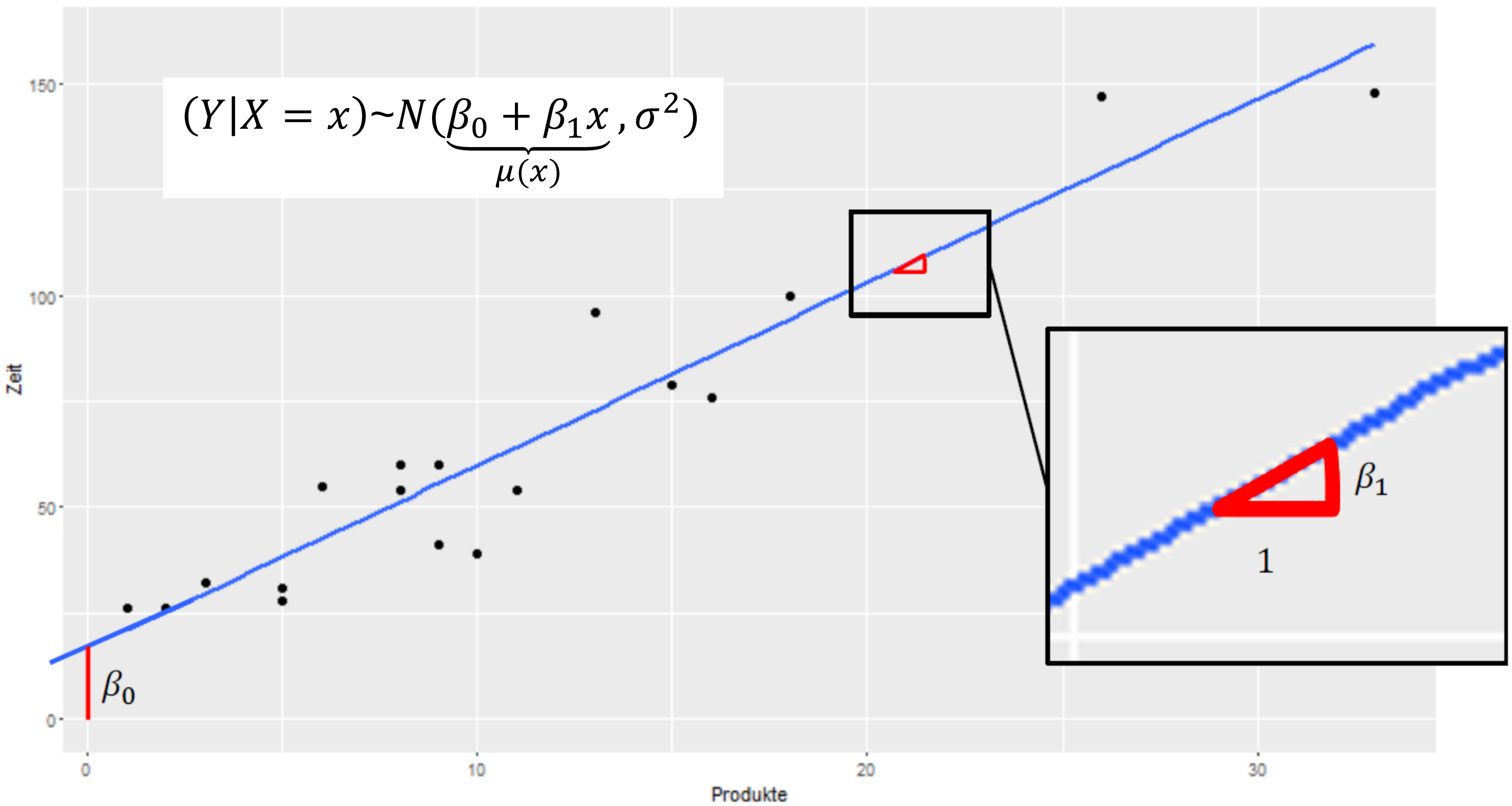
1. $(Y|X = x) \sim N(\underbrace{\beta_0 + \beta_1 x}_{\mu(x)}, \sigma^2)$

2. $Y = \beta_0 + \beta_1 x + \varepsilon$
   - $\varepsilon \sim N(0, \sigma^2)$

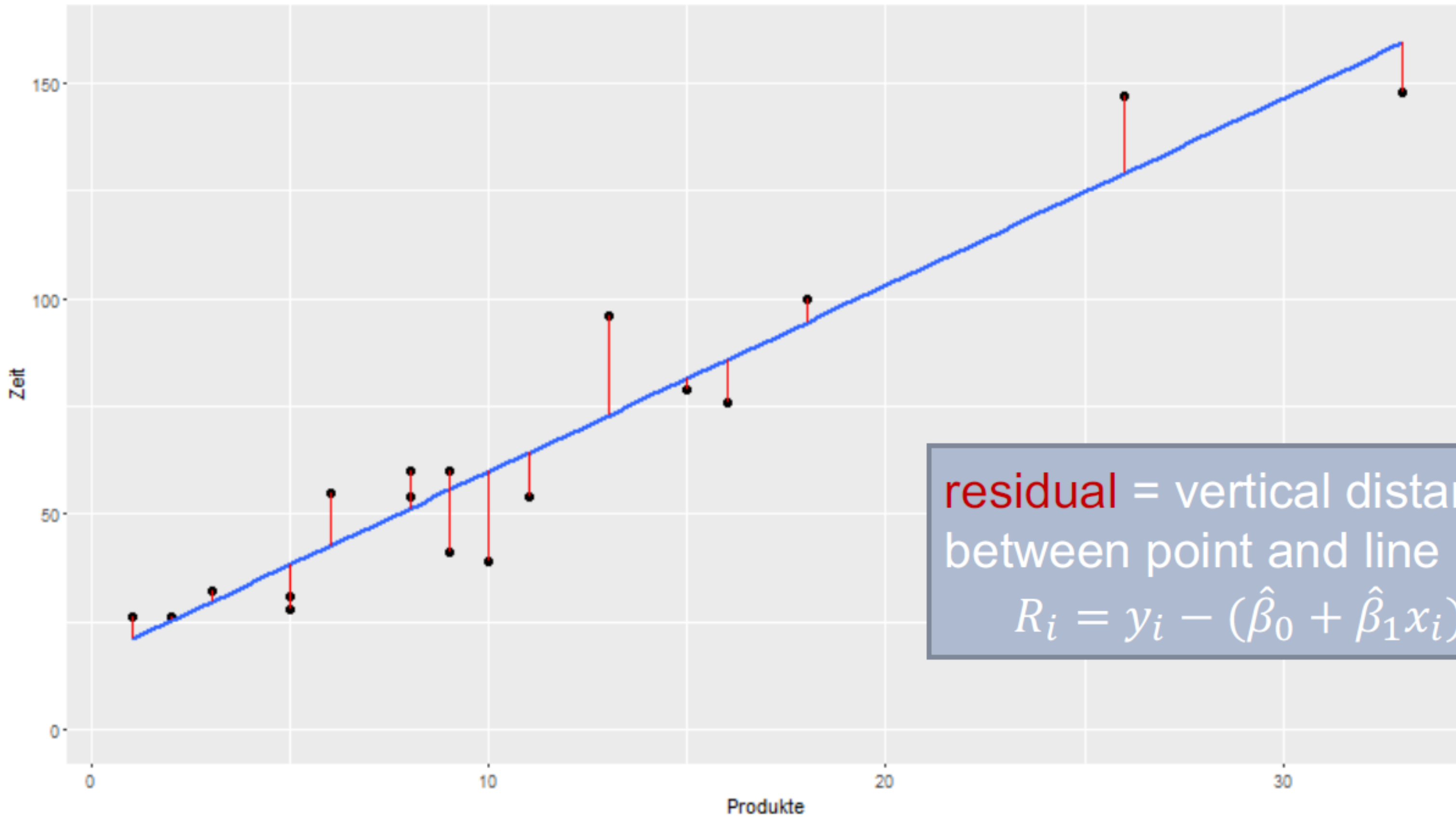
$$
\begin{aligned}
E(Y) &= E(\beta_0 + \beta_1 x + \varepsilon) \\
&= \beta_0 + \beta_1 x + E(\varepsilon) \\
&= \beta_0 + \beta_1 x \\
Var(Y) &= Var(\beta_0 + \beta_1 x + \varepsilon) \\
&= Var(\varepsilon) \\
&= \sigma^2
\end{aligned}
$$



Regression model



Residuals



residual = vertical distance between point and line

$$R_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Linear regression - setting the scene

Model for the condition probability distribution

CPD: $Y_{X_i} = (\mathrm{Y}|\mathrm{X}_i) \sim \mathrm{N}(\mu_{x_i}, \sigma^2)$

$\mathrm{Y}_x \in \mathbb{R}$, $\mu_x \in \mathbb{R}$

$\mathrm{y}_i = \beta_0 + \beta_1 \cdot x_{i1} + \varepsilon_i$

$\mathrm{E}\left(\mathrm{Y}_{X_i}\right) = \mu_{x_i} = (\mu|\mathrm{X}=\mathrm{x}_i) = \beta_0 + \beta_1 \cdot x_{i1}$

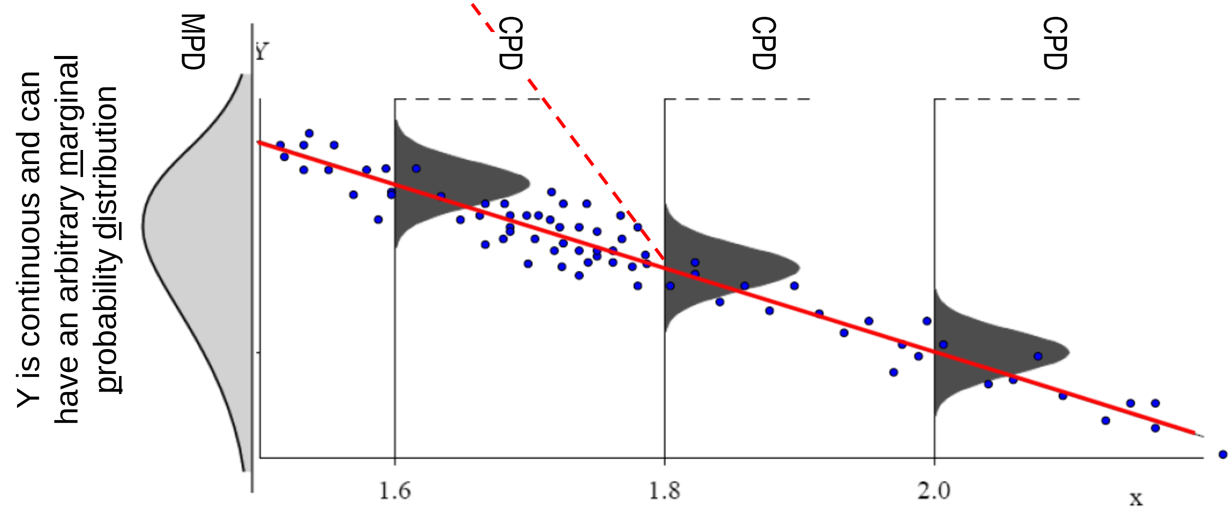
$\mathrm{Var}(\mathrm{Y}_{X_i}) = \mathrm{Var}(\mathrm{Y}|\mathrm{X}_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$

ε_i i.i.d. $\sim N(0, \sigma^2)$

identical independent distributed

$\mathrm{Y} \sim \mathrm{V}_{arbirary}^{contiuous}$

$(\mathrm{Y}|\mathrm{X}_i) \sim \mathrm{N}(\mu_{x_i}, \sigma^2)$



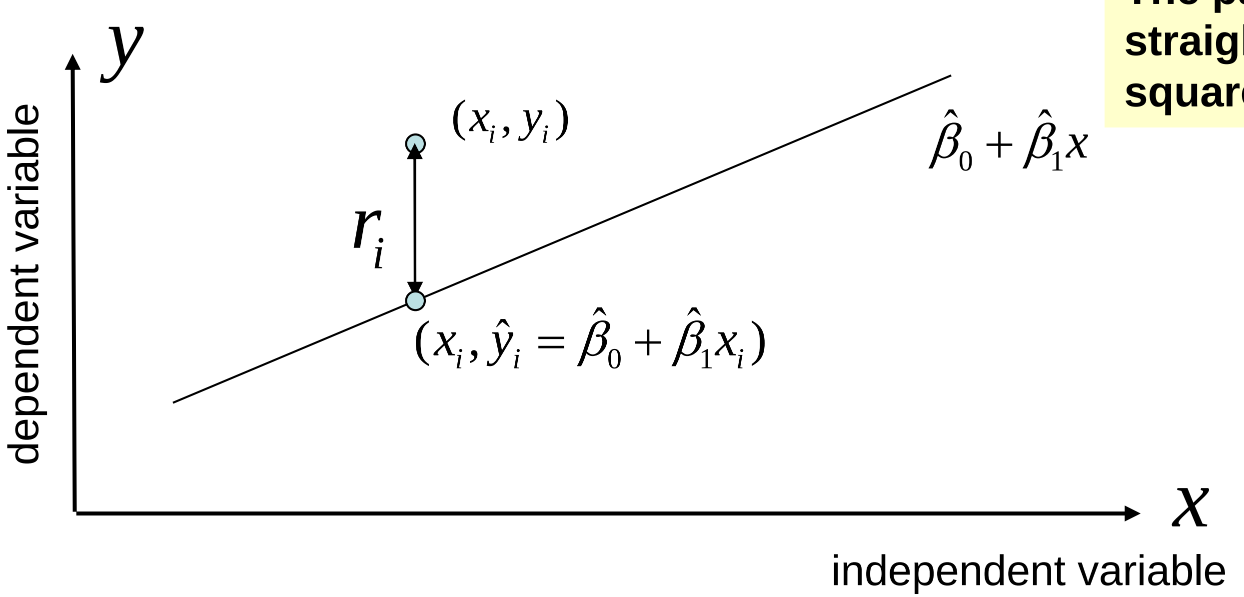
Regression model and residuals:

$$(Y|X=x) \sim N(\underbrace{\beta_0+\beta_1 x}_{\mu(x)}, \sigma^2)$$

The model has three parameters: $\beta_0, \beta_1, \sigma^2$

Illustration of the residuals

$$r_i = y_i - \hat{y}_i$$



The paradigm remains to fit a straight line such that the sum of squared residuals is minimized:

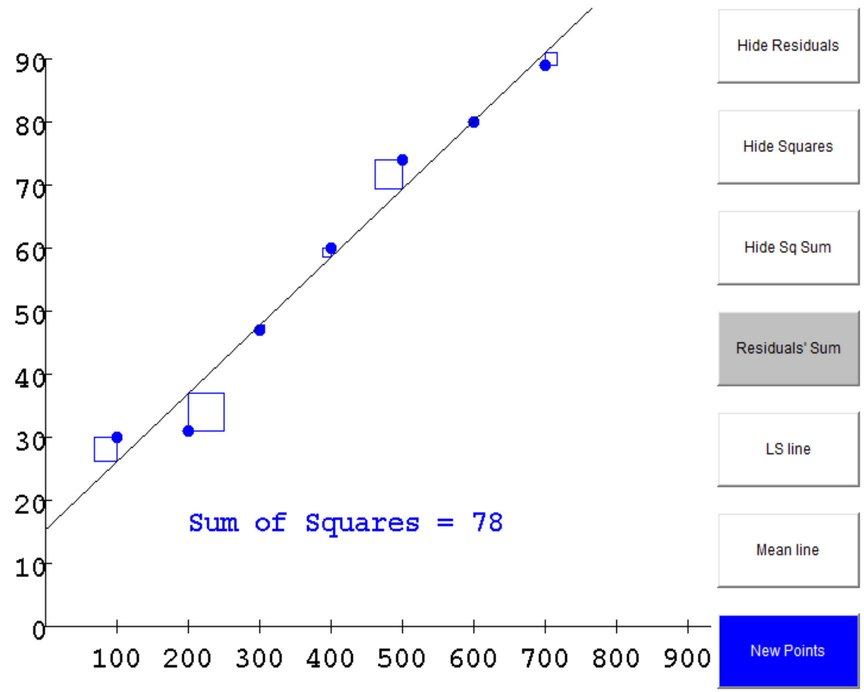
$$\sum_{i=1}^{n} r_i^2 = \min$$

Least Squares Fitting

We minimize the sum of squared residuals

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

Instructions for this demo are down below the graph.



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.

Least Squares: Estimation of the parameters

According to the least squares paradigm, the best fitting regression line is, i.e. the optimal coefficients are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \text{und} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

For a given set of data points $(x_i, y_i)_{i=1,\ldots,n}$, we can calculate the solution using the formulas above (or better we use R).

The numerical solution for our example "Tree Height":

$$\hat{\beta}_1 = -3.003, \;\; \hat{\beta}_0 = 28.723$$

→ **`lm(height ~ phvalue, data=treeheight)`**

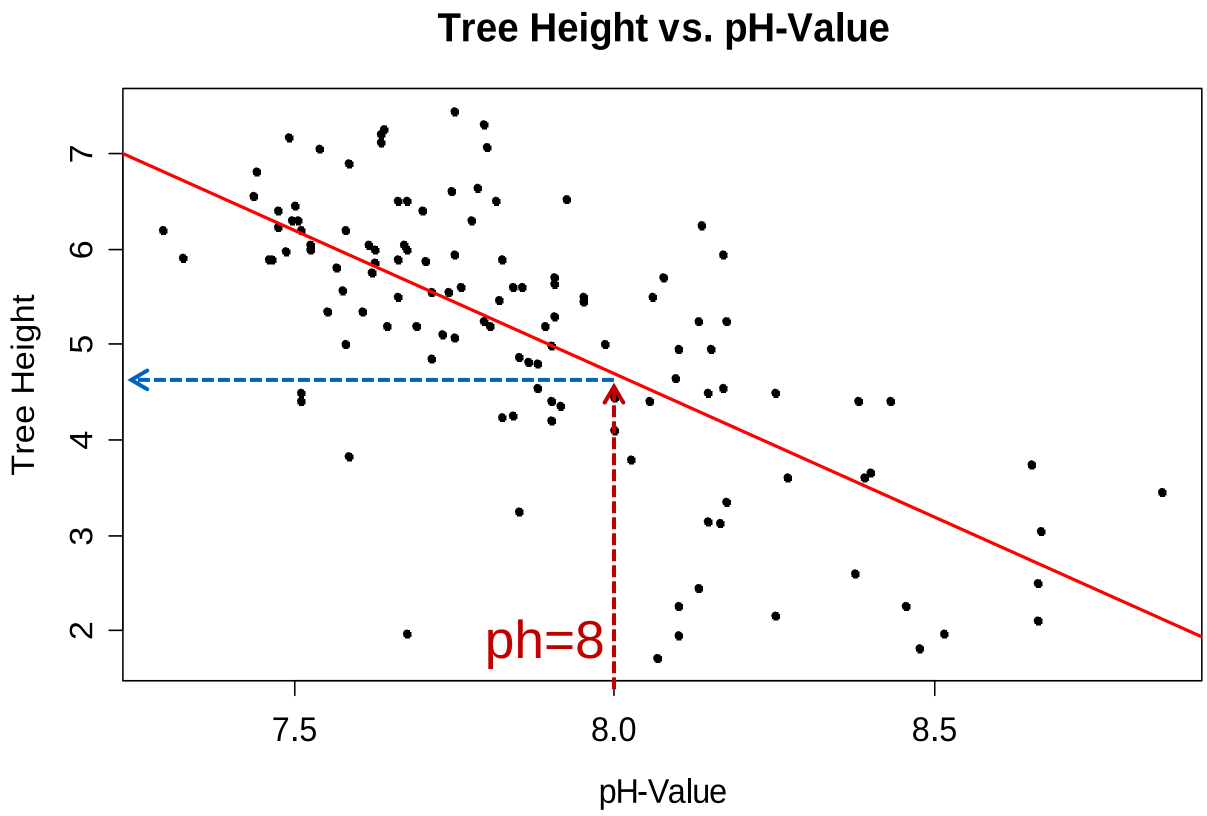
Estimation of the variance:

The variance can be estimated from the residuals.

$$\hat{\sigma}_E^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} r_i^2$$

The division by $n-(p+1)$ is for obtaining an unbiased estimator. Generally, n is the number of observation in the train dataset and p is the number of estimated regression coefficients.

Least Squares Regression Model



Prediction of the expected height (average of heights):

$$\text{height(ph)} = 28.7 - 3 \cdot \text{ph} \qquad \text{height}(8) = 28.7 - 24 = 4.7$$

Linear Regression for tree example in R

$$(Y|X=x) \sim N(\underbrace{\hat{\alpha}+\hat{\beta}\cdot x}_{\hat{\mu}(x)=\widehat{y(x)}}, \hat{\sigma}^2)$$

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

Intercept $\hat{\alpha}$

$se(\hat{\alpha})$

$t = \frac{\hat{\alpha} - \alpha_0}{se(\hat{\alpha})}$, $\alpha_0 = 0$

p_α-value

```
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 ***
```

$slope: \ \hat{\beta}$

$se(\hat{\beta})$

test value

p_β-value

$\hat{\sigma}_\varepsilon^2$

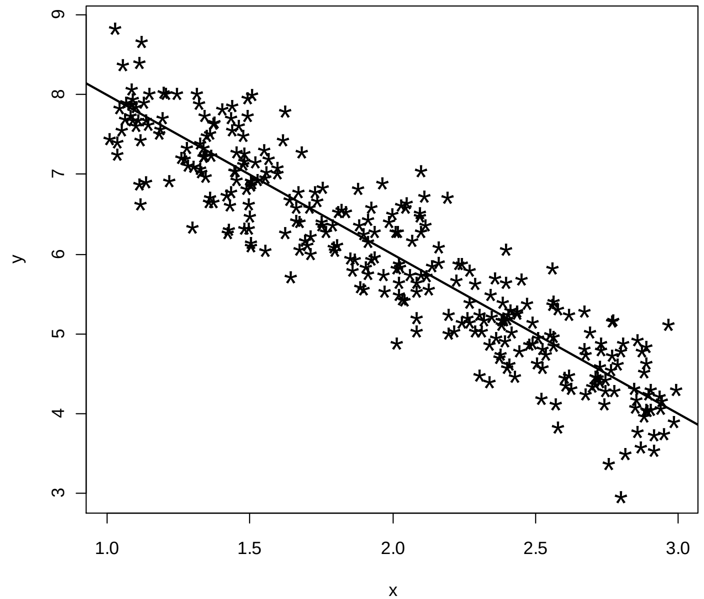
#datapoint-#estimated_parmeters

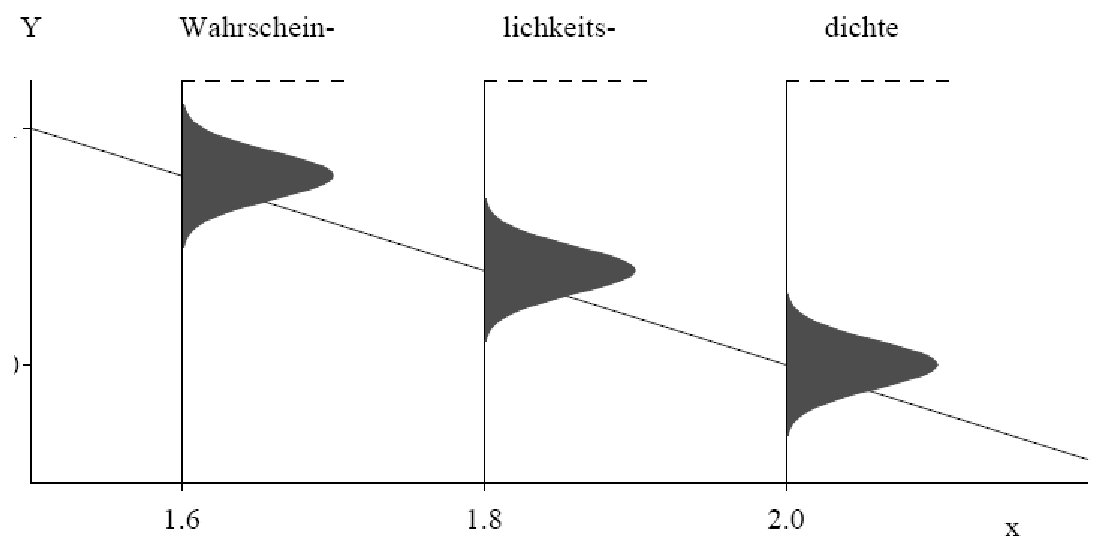
```
Residual stand. err.: 1.008 on 121 degrees of freedom
Multiple R-squared: 0.4797,
Adjusted R-squared: 0.4754
F-statistic: 111.5 on 1 and 121 DF,
p-value: < 2.2e-16
```

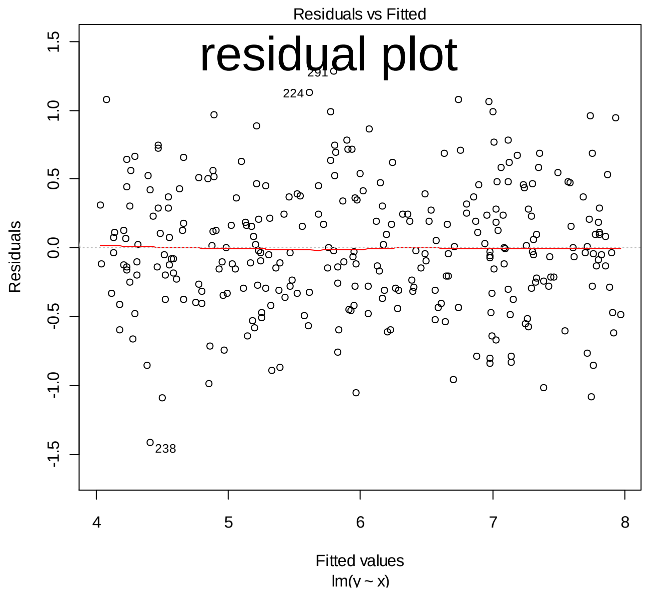
← R^2 (= corr^2 in case of 1 predictor)

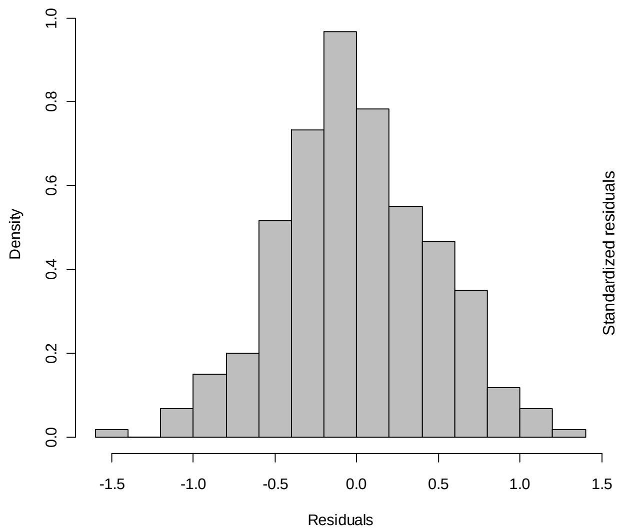
Global test for the model
(will see later)

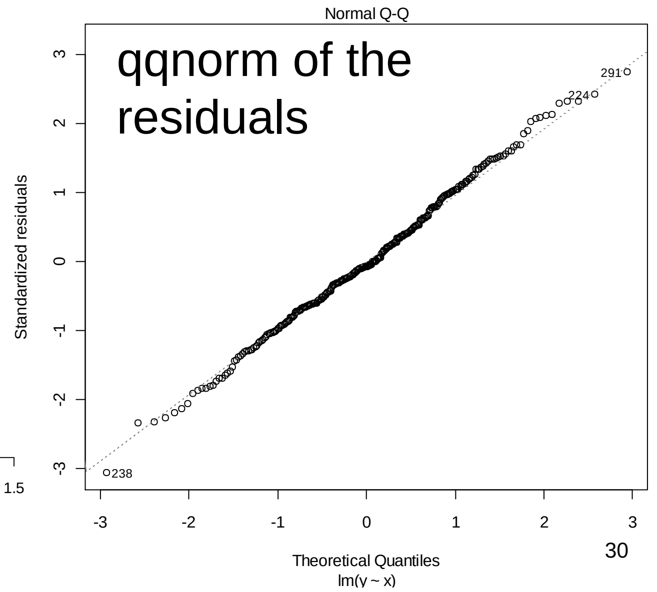
Check model assumption – are residuals iid $N(0, \sigma^2)$?





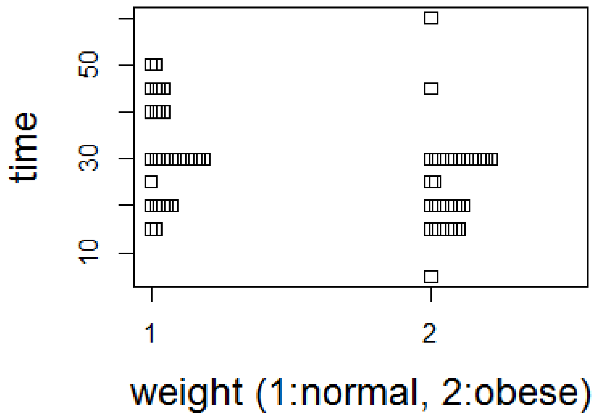


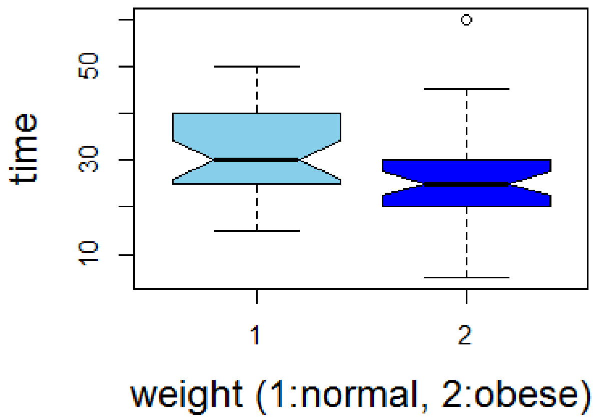




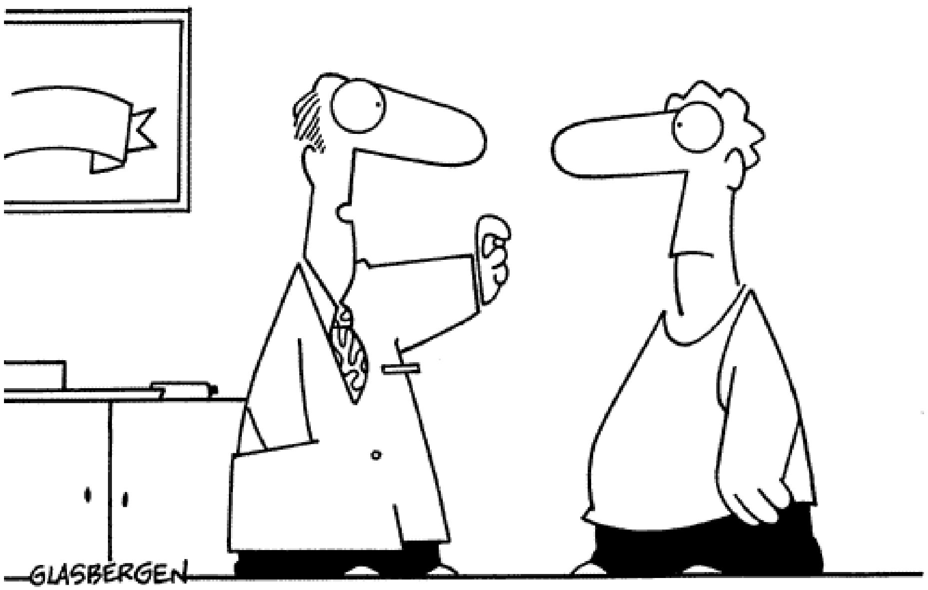
A binary variable as explanatory variables

Example with one factorial predictor
Do medical doctors spend less time with obese patients?





In an observational study it was measured how much time doctors spend with a patient.

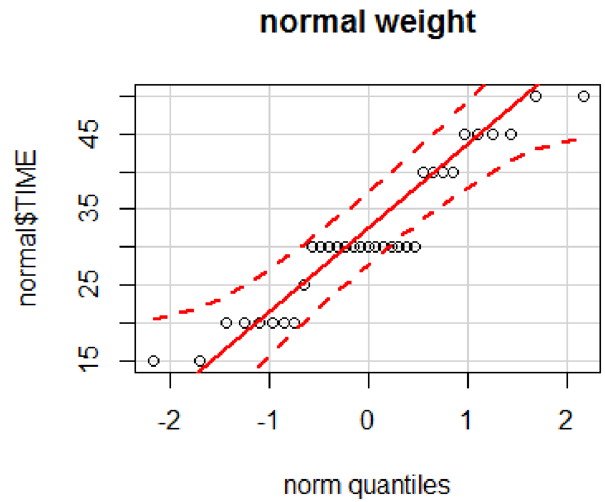


"To prevent a heart attack, take one aspirin every day. Take it out for a jog, then take it to the gym, then take it for a bike ride...."

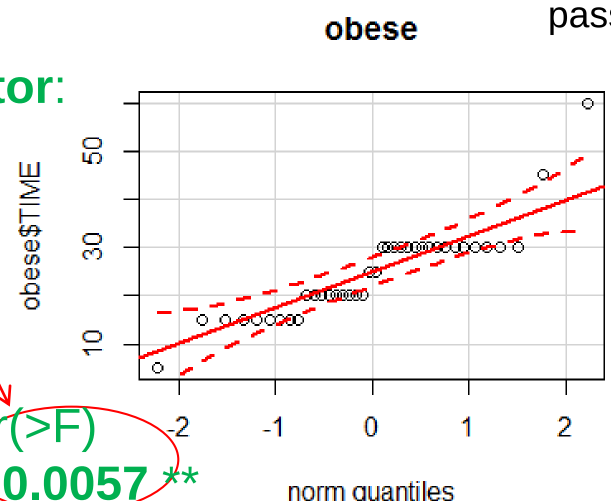
Do medical doctors spend less time with obese patients? How can we test this with linear regression and ANOVA?

t.test(TIME~WEIGHT, data=dat)

t = 2.9, df = 67, p-value = **0.0057**
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
2   11
sample estimates:
mean of x   mean of y
31   25



Normality check passed



do it by **regression with one factorial predictor**:

fit=**lm**(TIME~WEIGHT, data=dat)

anova(fit)

get anova-table from lm-object
Response: TIME

#	Df	Sum	Sq Mean	F value	Pr(>F)
# WEIGHT	1	776	776	8.16	**0.0057** **
# Residuals	69	6561	95		

An ANOVA with 1 factor with 2 levels is equivalent to a two-sample t-test.

Linear Regression with continuous and factorial predictors

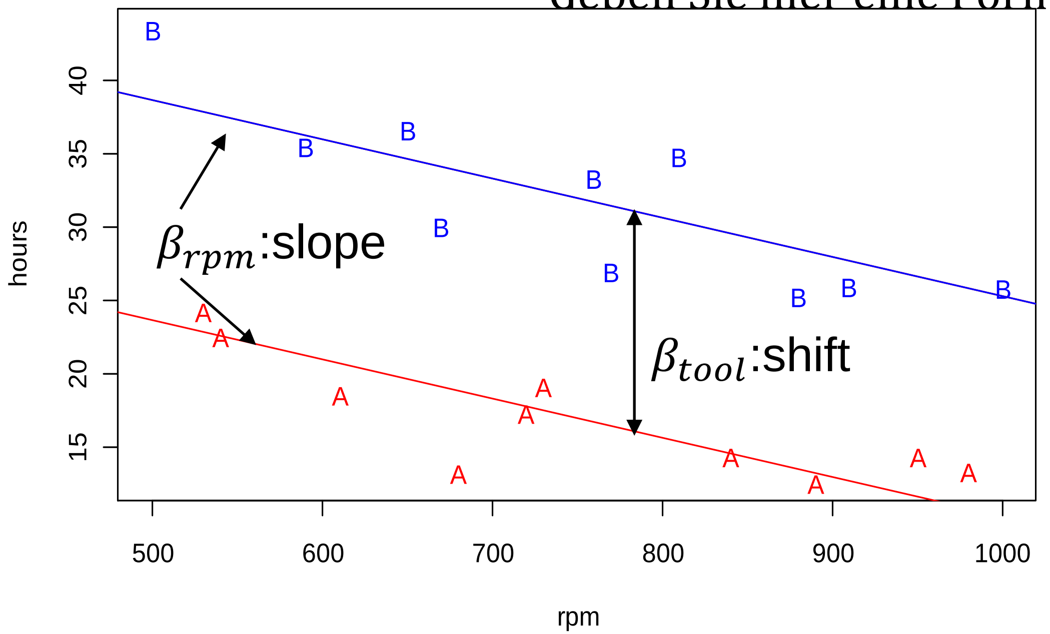
Output: **hours:** lifetime of a cutting tool

Predictor 1: **rpm**: speed of the machine in rpm (is a continuous variable)

Predictor 2: **tool:** tool type A or B (is a factor variable)

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

Geben Sie hier eine Formel ein.



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

Linear regression: interpretation of coefficient

$$(\hat{Y} | X = x) \sim N(\underbrace{\hat{\alpha} + \hat{\beta}_1 \cdot x_1 \underbrace{+ \hat{\beta}_2 \cdot x_2}}_{\hat{\mu}\,(x) = \hat{y}(x)}, \hat{\sigma}^2)$$

The coefficient β_1 of a continuous variable x_1 gives the change of the conditional mean of the outcome y, given the explanatory variable x_1 is increased by one unit and all other variables are hold constant.

The coefficient β_2 of a binary variable x_2 gives the change of the conditional mean of the outcome y, given the explanatory variable x_2 goes from the reference level (coded internally in R by 0) to the non-reference level (coded internally in R by 1) and all other variables are hold constant.

Summary

- Pearson correlation quantifies the strength of the linear associations
- Spearman rank correlation quantifies the strength of the monotone associations
- A simple linear regression models a conditional Gaussian distribution for the target variable Y in dependency on a single predictor X: $(Y|X = x) \sim N(\mu(x), \sigma^2)$ with $\mu(x) = \beta_0 + \beta_1 \cdot x$
- The residual of the i-th observation (x_i, y_i) is defined as $r_i = y_i - \mu(x_i)$
- Ordinary Leas Sqaure (OLS) estimates the parameters (β_0, β_1) as the values, for which the sum of the squared residuals is minimized
- The variance parameter σ is estimated from the residuals
- To check the model assumptions of a linear regression, we perform a residual analysis to check if the residuals r_i are iid $N(0, \sigma^2)$.