Biostatistics week 10

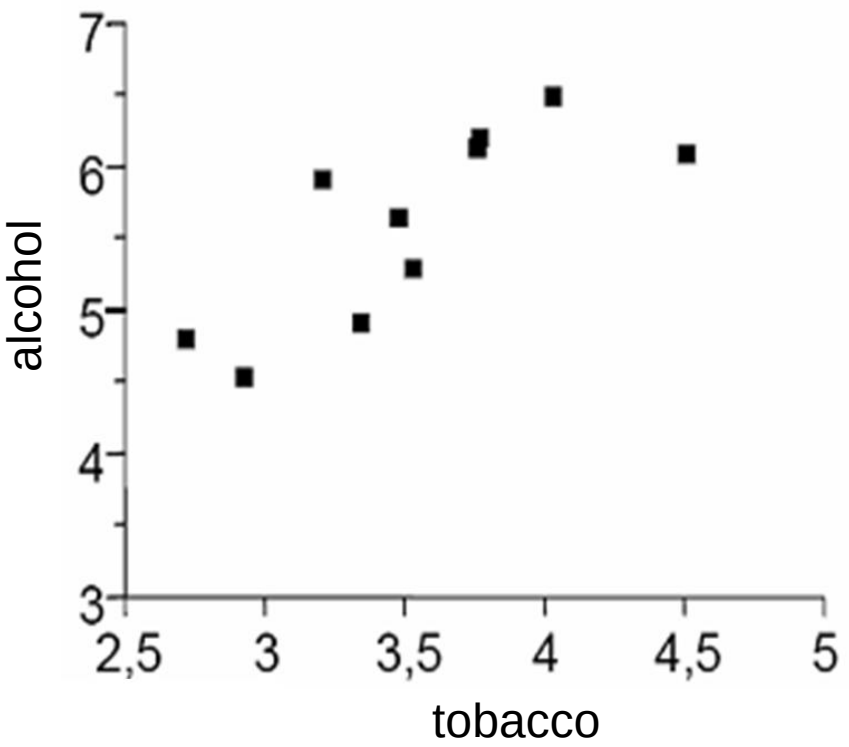
- Recap: correlation and parameter estimates in linear regression
- The origin of the term “regression”: Regression to the mean
- Coefficient of Determination R^2: unadjusted or adjusted
- Model and Variable selection with some warnings
- Linear regression with factor variables
  - interaction between a factor and a continuous predictor
  - t-Test or linear regression with a 2-level factor variable
  - One-way-ANOVA or linear regression with a factor variable
- Non-parametric tests for group comparison with >2 groups

Is there an association between 2 variables?

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

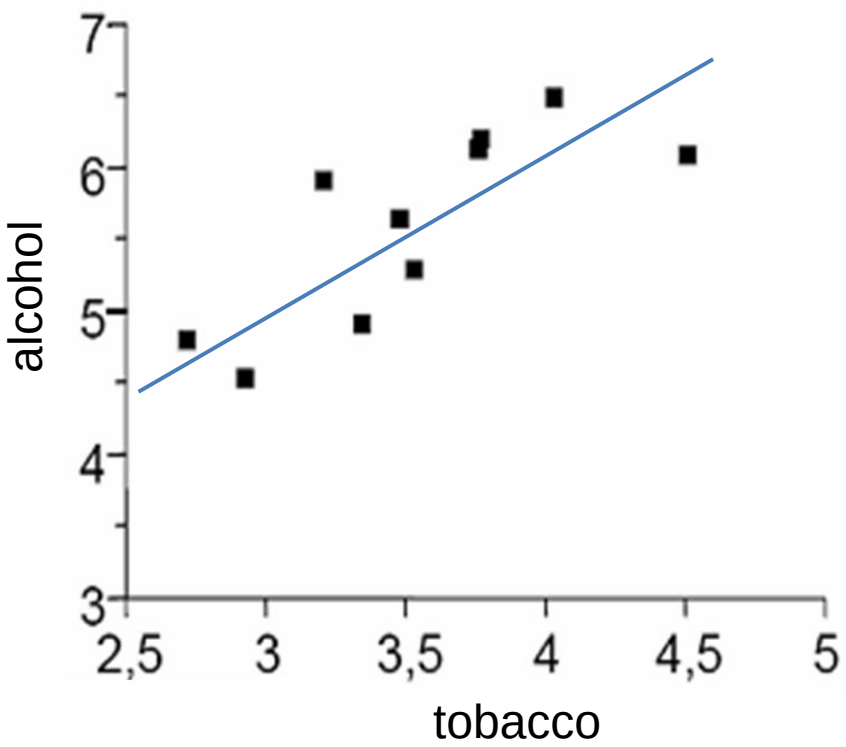
region	alcohol	tobacco
North	6,47	4,03
Yorkshire	6,13	3,76
Northeast	6,19	3,77
East Midlands	4,89	3,34
West Midlands	5,63	3,47
East Anglia	4,52	2,92
Southeast	5,89	3,2
Southwest	4,79	2,71
Wales	5,27	3,53
Scotland	6,08	4,51

$cor_{Pearson}$=0.83



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

Fit a linear regression model



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

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:

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

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

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:

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

Slope given by $\Delta y/\Delta x = sd(y)/sd(x)$ which gets shrinked in case of non deterministic relationships

Intercept

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

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

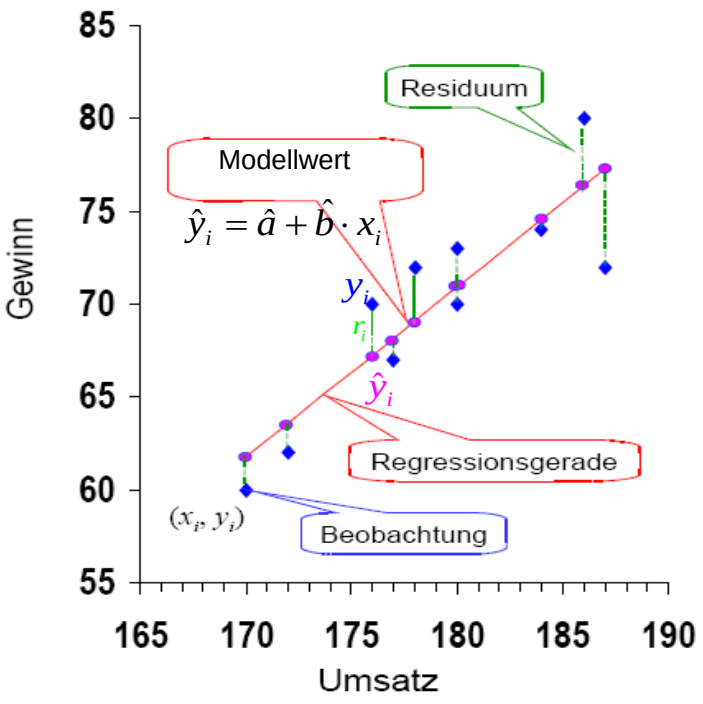
Standard-deviation of X and Y

Steigung:

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

y-Achsenabschnitt:

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



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

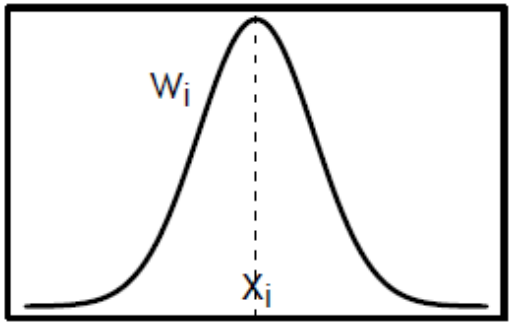
$$= \frac{1}{n-2} \sum_{i=1}^{n} (y_i - (\hat{a} + \hat{b} \cdot x_i))^2$$

Side track: Errors in explanatory variables

In regression we assume that the explanatory variables are fixed and error-free.

However, in some situations we should assume that the error-free x cannot be observed and instead we observe an error-prone value w.

Classical ME model



$$W_i = x_i + U_i$$

$$U_i \sim N(0, \sigma_u^2)$$

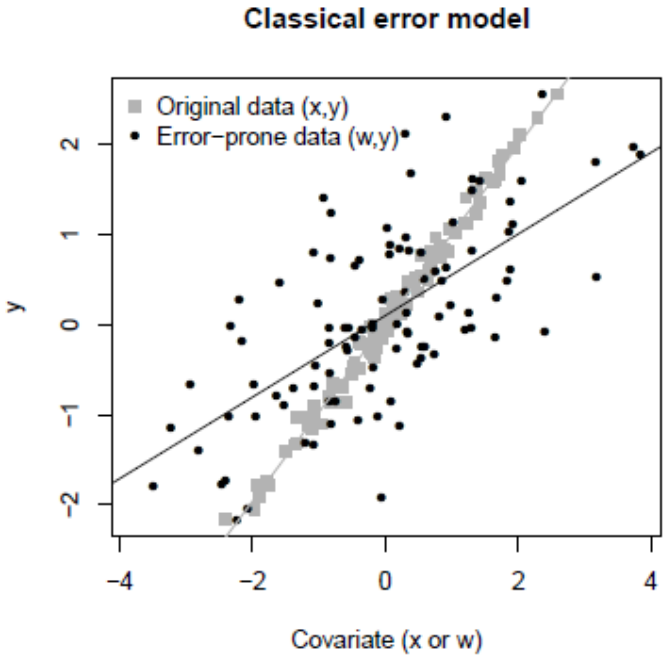
Side track: Errors in explanatory variables ctd.

We show know that the slope is underestimated if the explanatory variable is not error free:

true model: $y = \beta_0 + \beta_x \cdot x + \varepsilon$

error model: $w = x + u \quad , \quad u \perp y, \varepsilon$

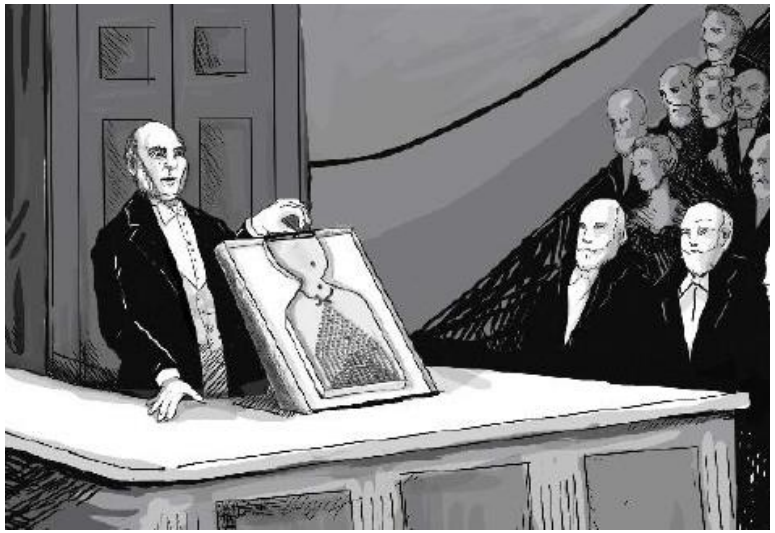
naive model: $y = \beta_0^* + \beta_x^* \cdot w + \varepsilon$



estimate with true predictor: $\hat{b} = r_{xy} \cdot \frac{s_y}{s_x}$

naive estimate: $\hat{b}^* = r_{(x+u)y} \cdot \frac{s_y}{s_{(x+u)}} = r_{xy} \cdot \frac{s_y}{s_x + s_u} < \hat{b}$

Galton on the search for causality



Galton in 1877 at the Friday Evening Discourse at the Royal Institution of Great Britain in London.

Francis Galton (first cousin of Charles Darwin) was interested to **explain** how traits like “intelligence” or “height” is passed from generation to generation.

Galton presented the “quincunx” (Galton nailboard) as causal model for the inheritance.

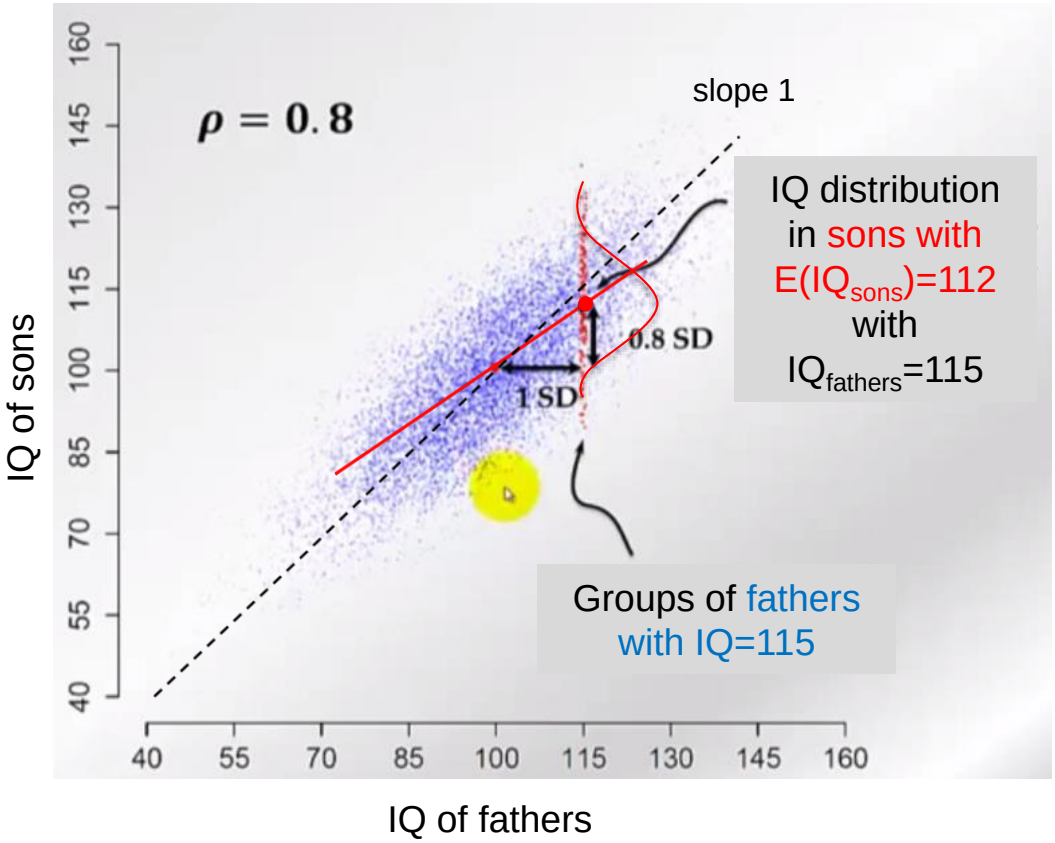
Balls “inherit” their position in the quincunx in the same way that humans inherit their stature or intelligence.

The stability of the observed spread of traits in a population over many generations contradicted the model and puzzled Galton for years.

Image credits: “The Book of Why”

Galton's discovery of the regression line

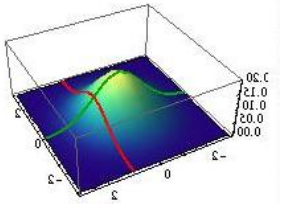
Remark: Correlation of IQs of parents and children is only 0.42 https://en.wikipedia.org/wiki/Heritability_of_IQ



$$X1 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

$$X2 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & \mathrm{cov}(X1,X2) \\ \mathrm{cov}(X1,X2) & 15^2 \end{pmatrix}\right)$$

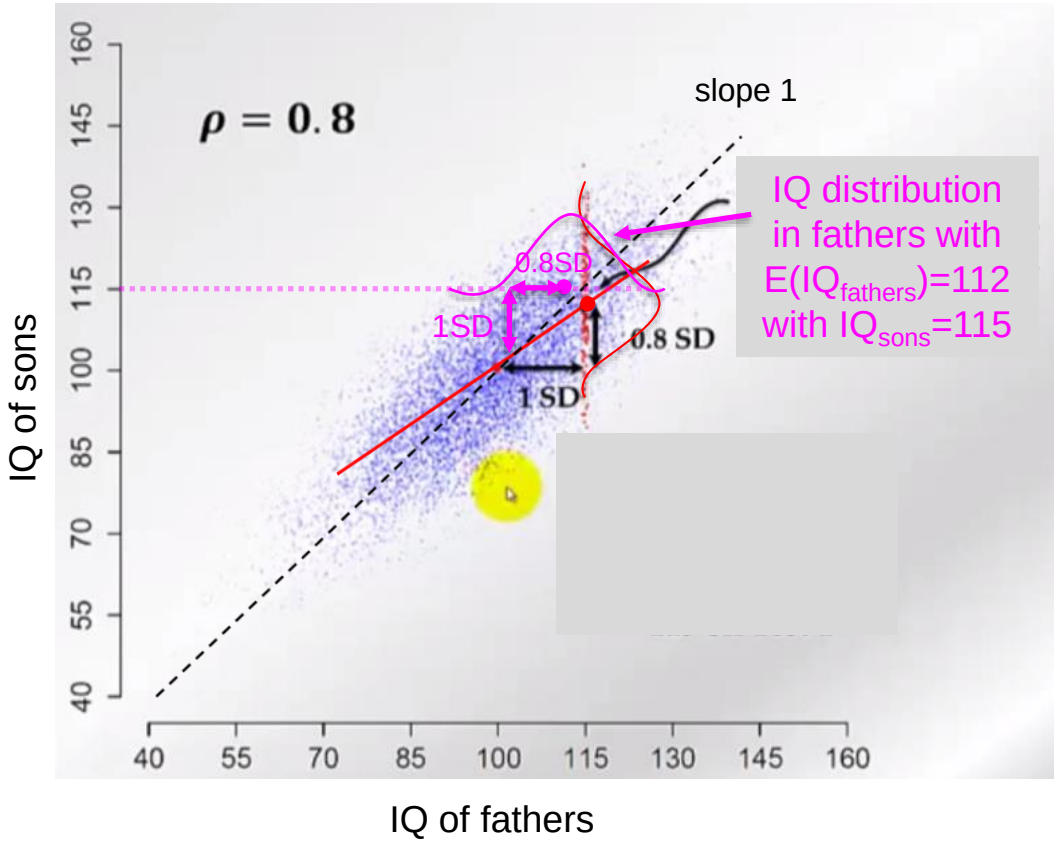


For each group of father with fixed IQ, the mean IQ of their sons is closer to the overall mean IQ (100) -> Galton aimed for a causal explanation.

All these predicted $E(IQ_{son})$ fall on a "regression line" with slope<1.

Image credits (changed): https://www.youtube.com/watch?v=aLv5cerjV0c

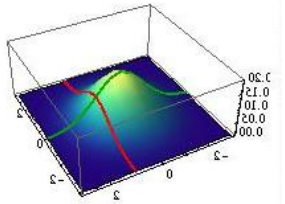
Galton's discovery of the regression to the mean phenomena



$$X1 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

$$X2 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

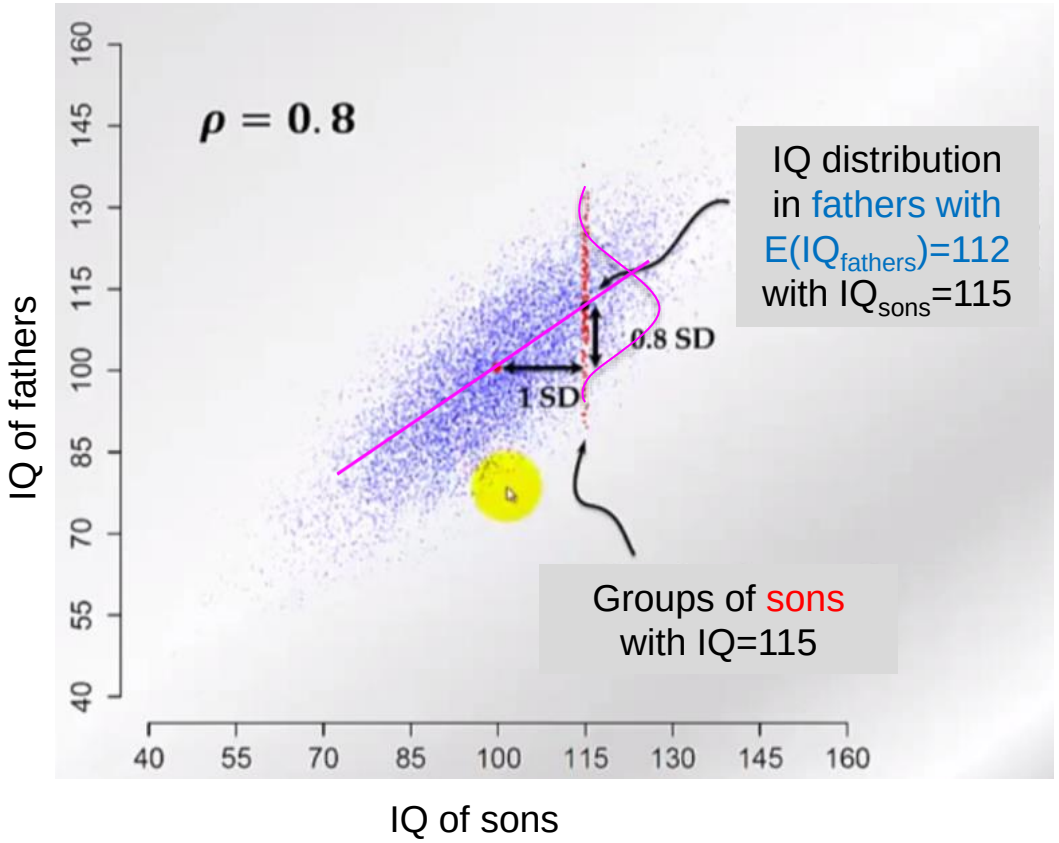
$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & \text{cov}(X1,X2) \\ \text{cov}(X1,X2) & 15^2 \end{pmatrix}\right)$$



Also the mean of all fathers who have a son with IQ=115 is only 112.

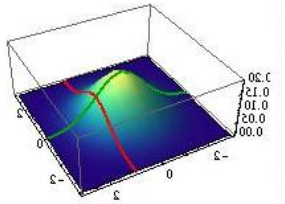
Image credits (changed): https://www.youtube.com/watch?v=aLv5cerjV0c

Galton's discovery of the regression to the mean phenomena



$$X1 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$
$$X2 \sim N\left(\mu_1 = 100,\ \sigma_1^2 = 15^2\right)$$

$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & \text{cov}(X1,X2) \\ \text{cov}(X1,X2) & 15^2 \end{pmatrix}\right)$$

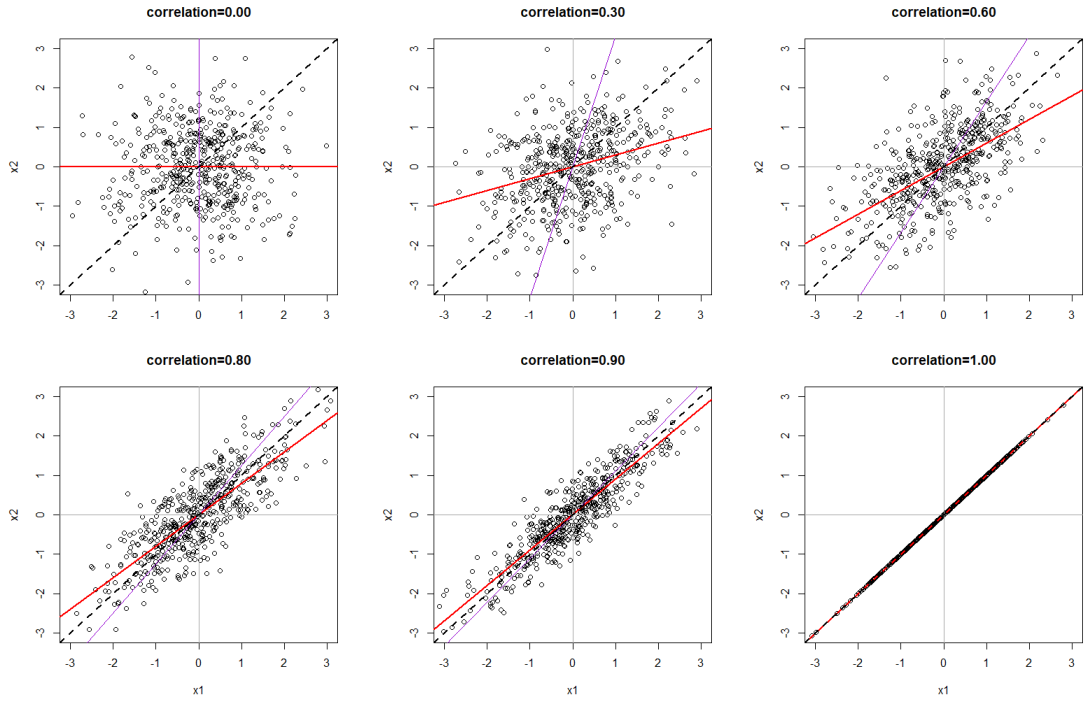


After switching the role of sons's IQ and father's IQ, we again see that $E(IQ_{fathers})$ fall on the regression line with the same slope <1.

There is no causality in this plot -> causal thinking seemed unreasonable.

Image credits (changed): https://www.youtube.com/watch?v=aLv5cerjV0c

Pearson's mathematical definition of correlation unmasks "regression to the mean" as statistical phenomena



The correlation c of a bivariate Normal distributed pair of random variables are given by the slope of the regression line after standardization!

c quantifies strength of linear relationship and is only 1 in case of deterministic relationship.

After standardization of the RV:

$$X1 \sim N\left(\mu_1 = 0,\ \sigma_1^2 = 1^2\right)$$
$$X2 \sim N\left(\mu_2 = 0,\ \sigma_2^2 = 1^2\right)$$

$$\begin{pmatrix} X1 \\ X2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{X1}^2 = 1 & c \\ c & \sigma_{X2}^2 = 1 \end{pmatrix}\right)$$

Regression line equation:

$$\hat{X}_2 = E\left(X_2 \mid X_1\right) = \beta_0 + \beta_1 \cdot X_1$$

$$\beta_1 = c \cdot \frac{\sigma_2}{\sigma_1} \overset{\text{stand.}}{=} c$$

β_1 quantifies regression to the mean

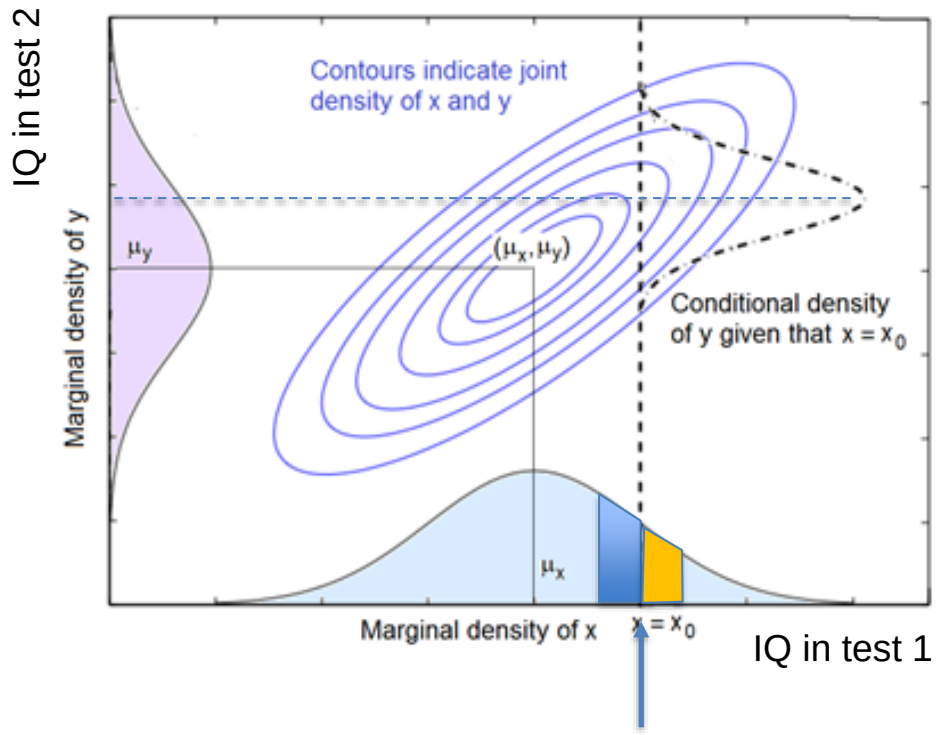
$$\beta_0 = \mu_2 - \beta_1 \cdot \mu_1 \overset{\text{stand.}}{=} 0$$

$$c = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)\cdot(x_{i2} - \bar{x}_2)}{\mathrm{sd}(x_1)\cdot\mathrm{sd}(x_2)}$$

Intuitive explanation of "regression to the mean"

IQ test result (at both time points) = true IQ + luck or bad luck

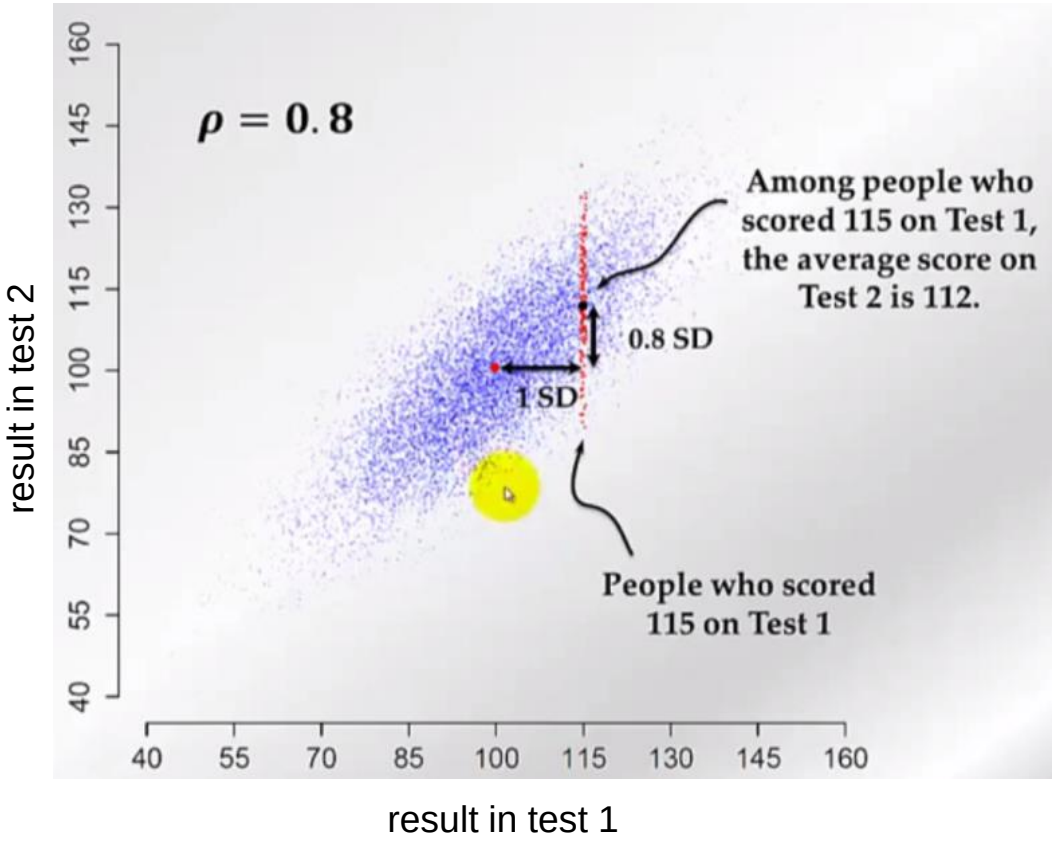
Not reproducible in second test



To get this test result, a person might

- have truly this high IQ (this are some people)
- have a lower true IQ (**many people** have a lower IQ) but **had luck**
- have a higher true IQ (**fewer people** have a higher IQ) but **had bad luck**

Regression to the mean occurs in all test-retest situations



Retesting a extreme group (w/o intervention in between) in a second test leads in average to a results that are closer to the overall-mean -> to assess experimentally the effect of an intervention also a control group is needed!

Regression to the mean gets easily forgotten

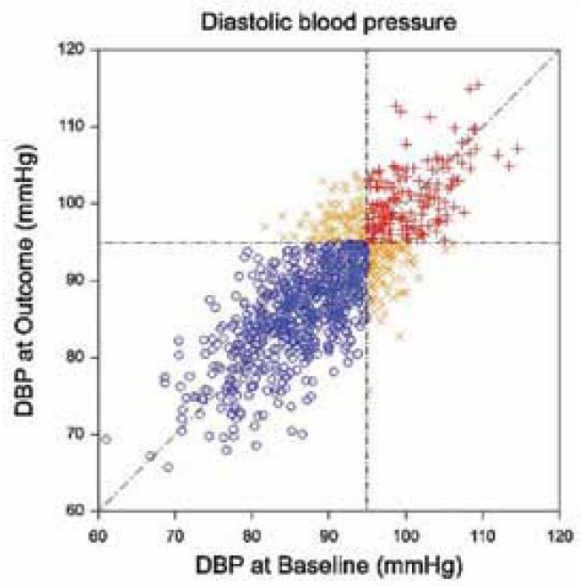
Three things that every medical writer should know about statistics

by Stephen Senn

Introduction

The joke goes that there are three kind of statistician: those who can count and those who can't. Therefore, readers of the *Write Stuff* will forgive me, I hope, if I end up writing about more than three things. It should be obvious, in that case, as to which sort of statistician I am. There are, of course, many more things than three that every medical writer should know about statistics because there are many things about statistics that anybody working in drug development should know and medical writers are in the unenviable position of having to know about everything. However, everybody has to start somewhere and three is a number with a great tradition. The three things I am going to write about are *regression to the mean [1]*, the *error of the transposed conditional [2]* and *individual response [3]*. The first is a widespread phenomenon that has a powerful influence on the way that results appear to us, the second is a pernicious fallacy and the third is a sort of Holy Grail-cum-wild goose chase that is responsible for leading many a researcher astray.

Figure 1 Simulated results at baseline and outcome for diastolic blood pressure (mmHg) for 1000 individuals in a population.



Regression to the mean

Regression to the mean is the tendency for members of a population who have been selected because they are extreme to be less extreme when measured again [4, 5]. Be-

Now consider a plot of a subset of the individuals, namely those who are 'hypertensive' on at least one occasion. This plot is given in figure 2. Just as was the case in figure 1 there is no essential difference as to whether we look at

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.

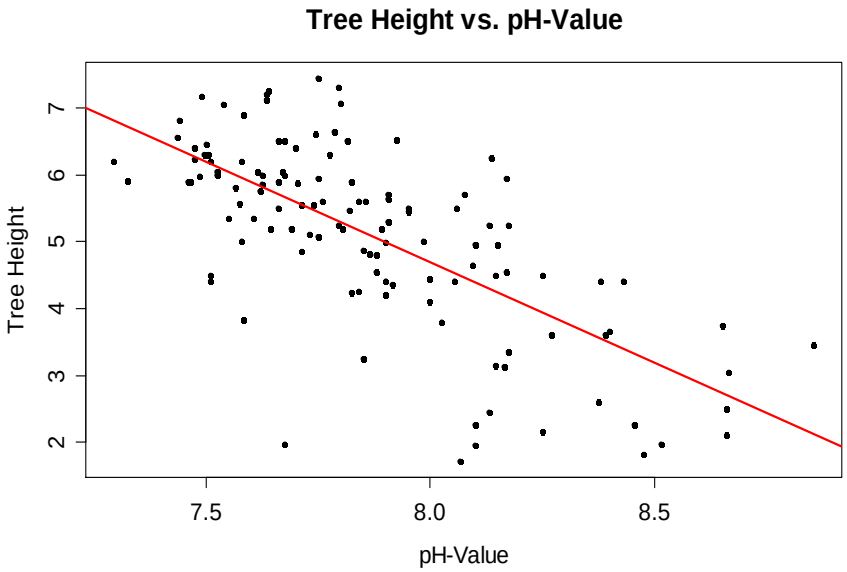
In case of regression x is the predictor variable (independent variable, explanatory variable), which is known precisely and has no error term

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.

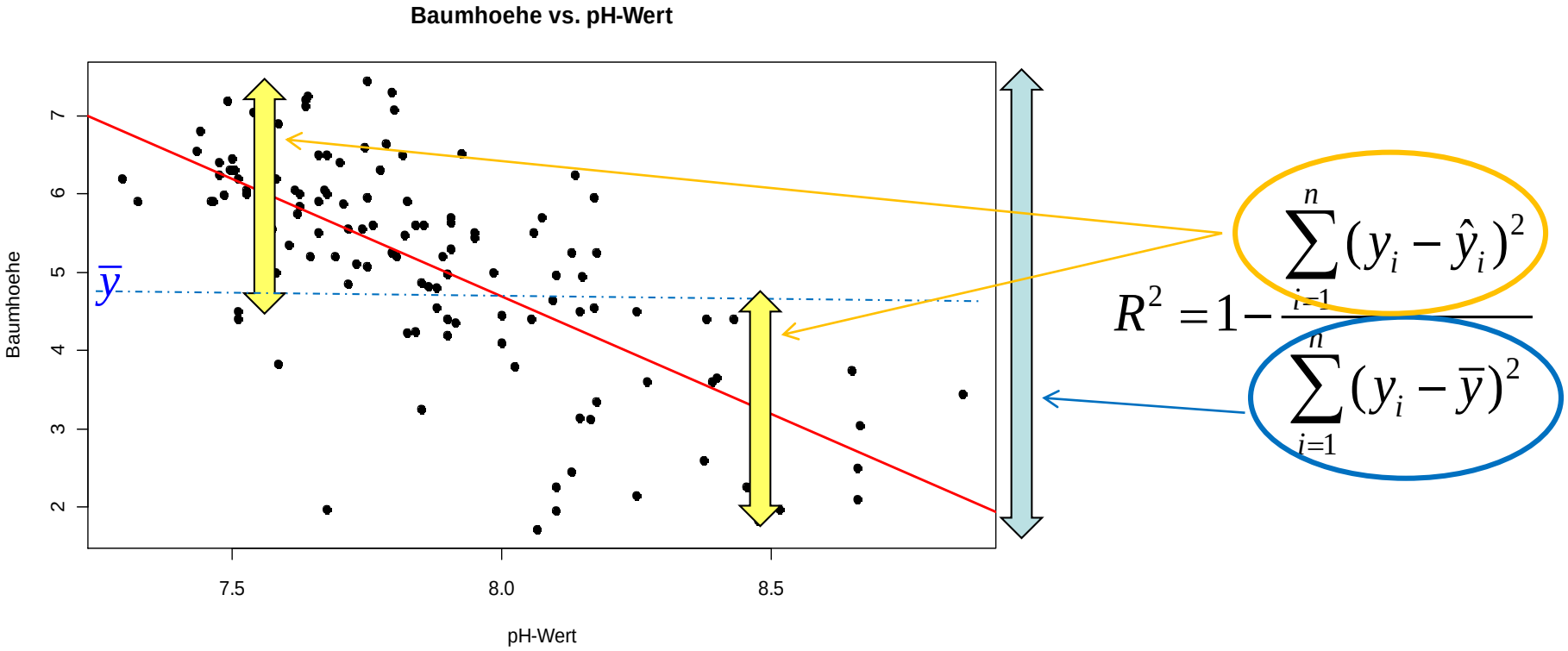
Investigate ph effect on height of trees

```
> 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?



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* R^2 *,named Coefficient of Determination* is often used as performance meassure. :

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

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.

adjR²: 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!**

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

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)

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

against the alternative (we need at least 1 predictor)

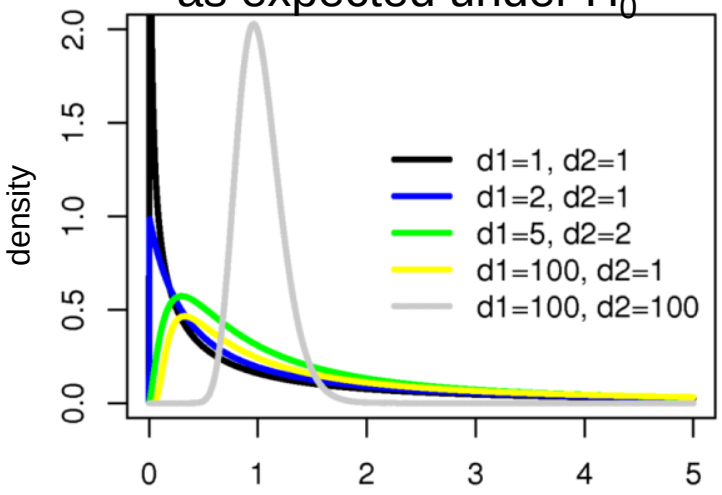
for at least one j in 1,…, p

$$H_A : \beta_j \neq 0$$

The test statistic is:

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

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:

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

```
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

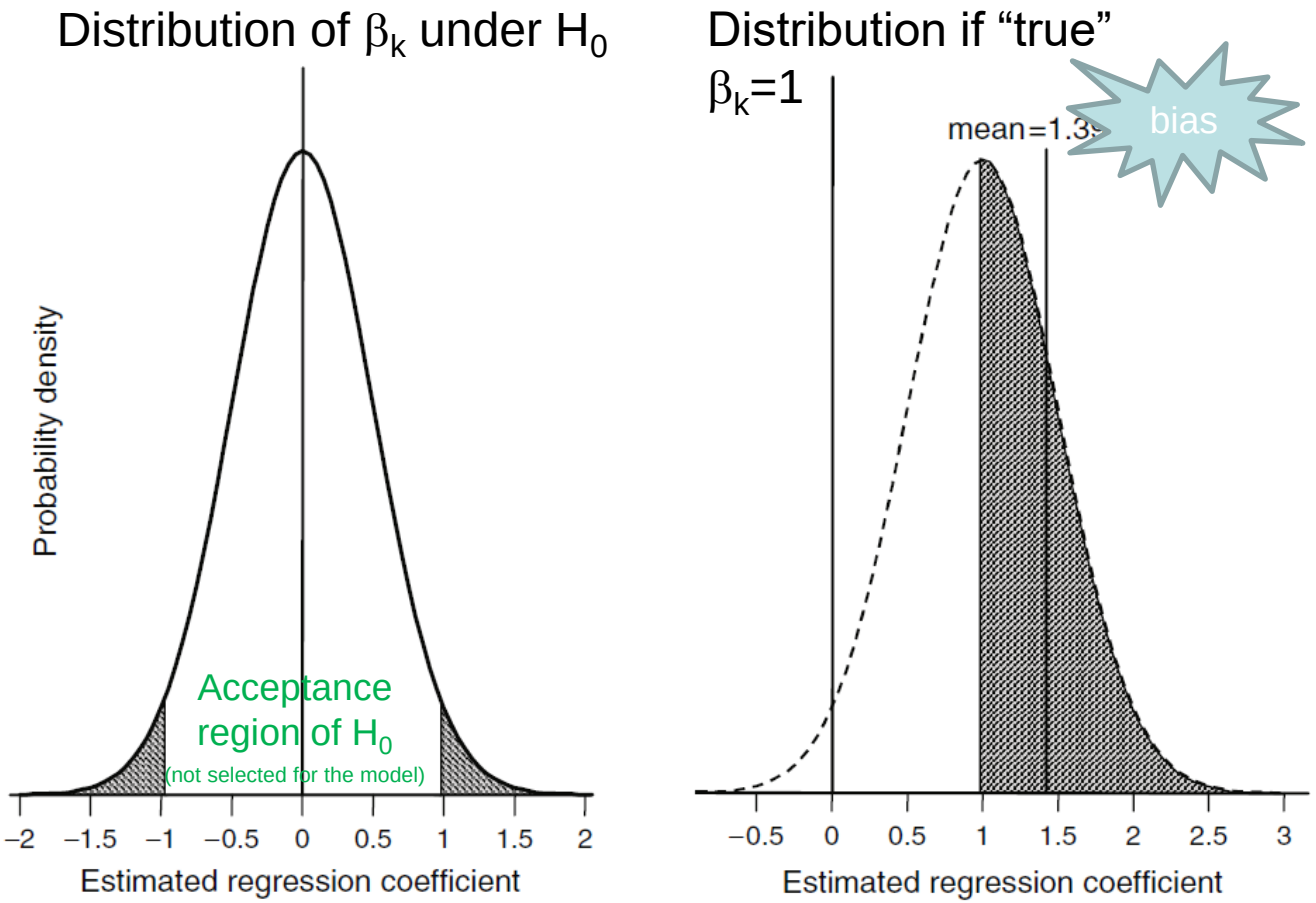


Fig. 5.5 Illustration of testimation bias. In case of a noise variable, the average of estimated regression coefficients is zero, and 2.5% of the coefficients is below −0.98 (1.96 × SE of 0.5), and 2.5% of the coefficients is larger than +0.98 (1.96 × SE of 0.5). In case of a true coefficient of 1, the 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

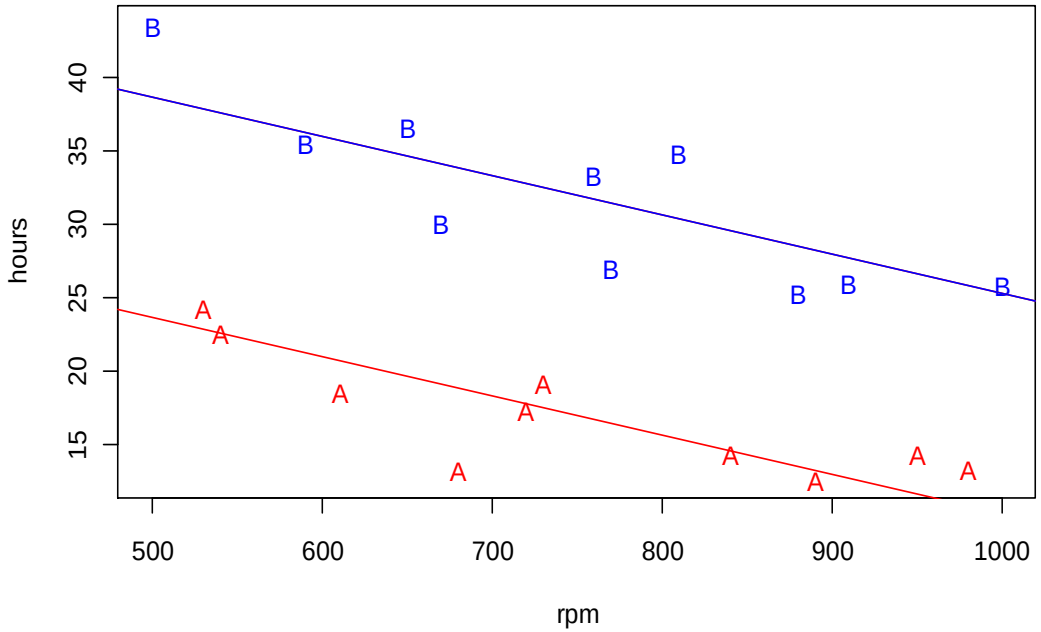
Linear Regression with continuous and factorial predictors

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)
```

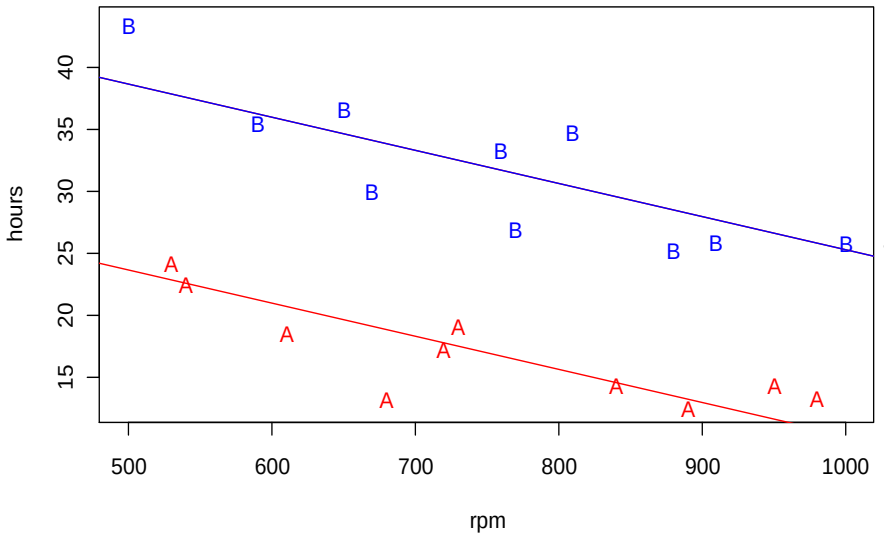


We have an additive model: the difference between the tools is a shift.

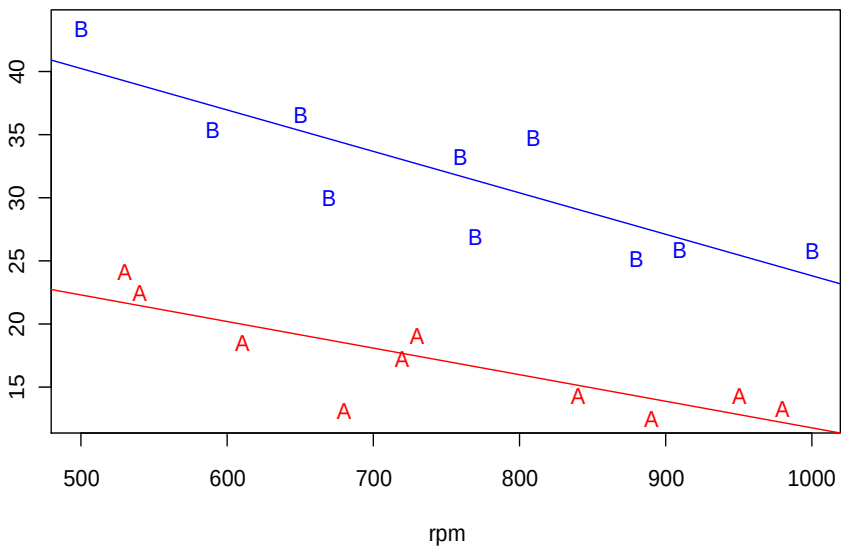
What does interaction mean?

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?

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

toolB (toolB=1):

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

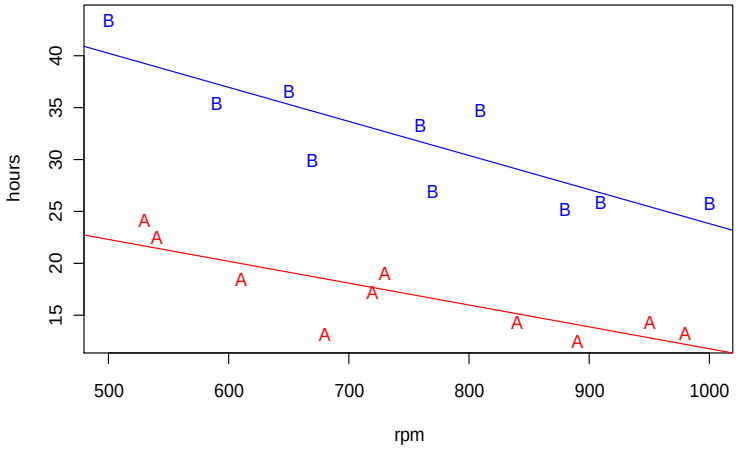
$$\text{hour} = 56.9 - 0.03 \cdot \text{rpm}$$

toolA (toolB=0):

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

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

Interaction is allowed

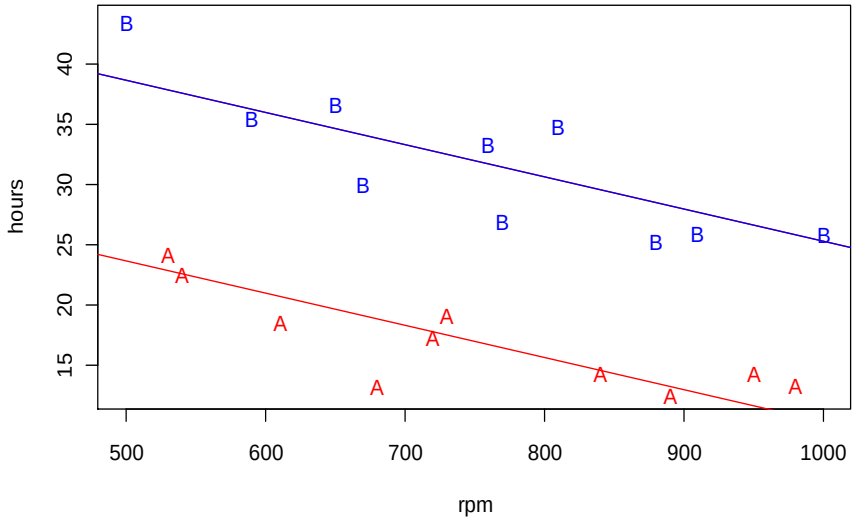


In case of interaction, the slope of the predictor “rpm” changes for different levels of the second predictor “tool” – also the intercept is changing for the two tools.

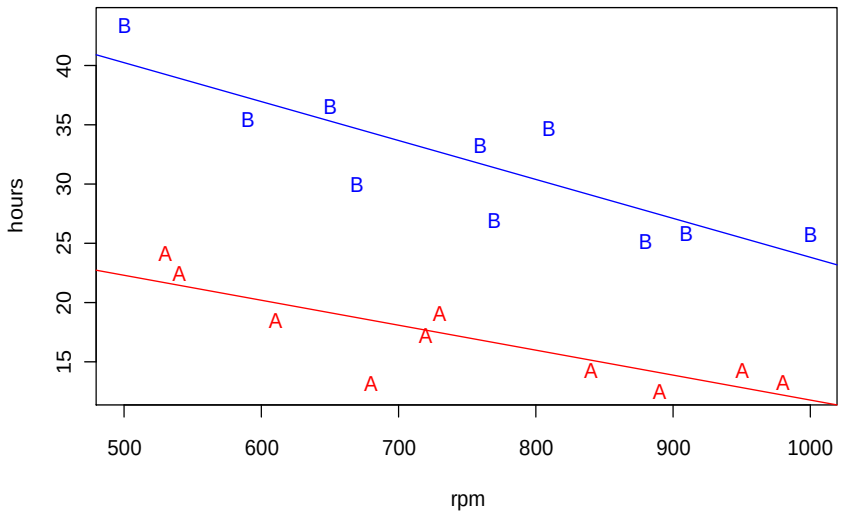
Remark: In case of interaction between two continuous predictors, slope (and intercept) of one predictor changes continuously with a continuous changing value of the other predictor and vice versa.

Do we need the complex model with the interaction?

Do not allow for interaction



Interaction is allowed

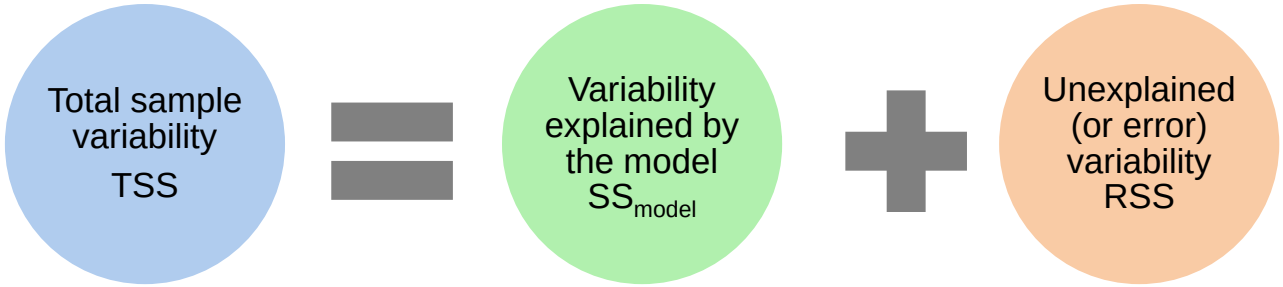



```
fit1=lm(hours ~ rpm + tool,
        data=my.dat)
```

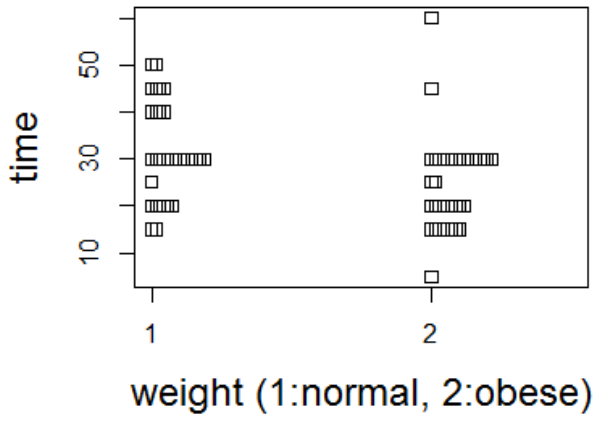
```
fit2=lm(hours ~ rpm * tool,
        data=my.dat)
```

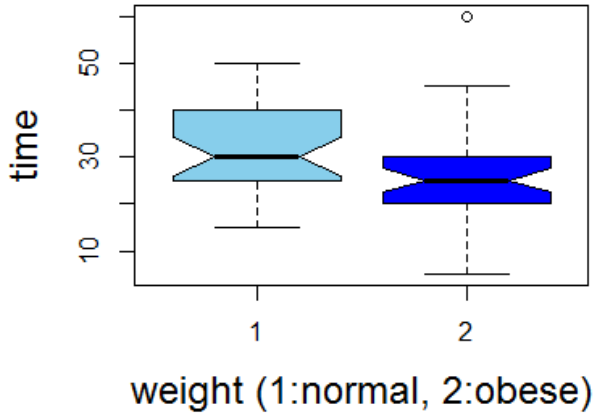
```
anova(fit2,fit1)
# p>5%, therefore interaction is not needed
```

ANalysis Of Variance (ANOVA)
= linear regression with factor variables

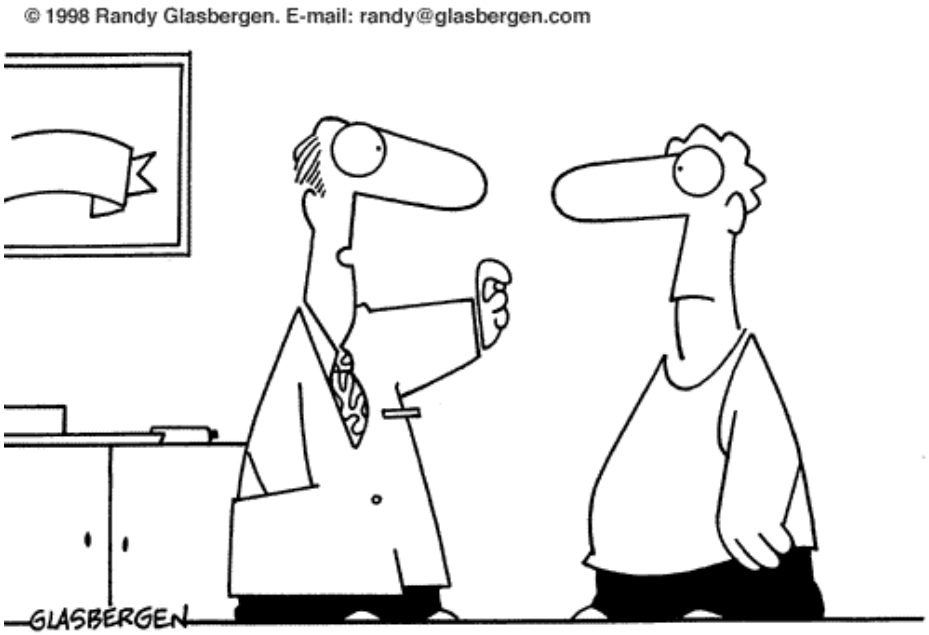


Example with one factorial predictor
Do medical doctors spend less time with obese patients?



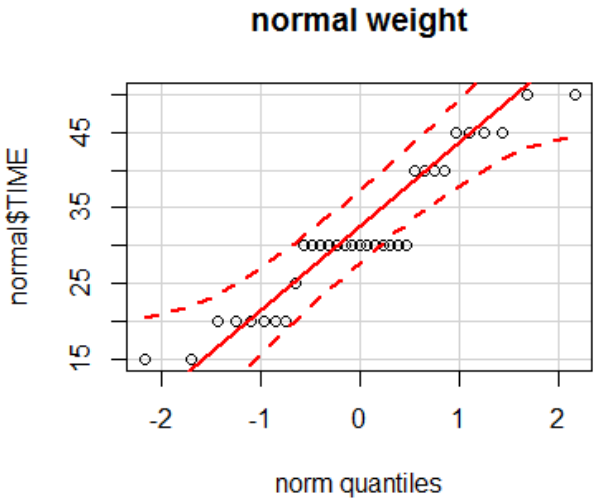


In an observational study it was measured how much time doctors spend with a patient.

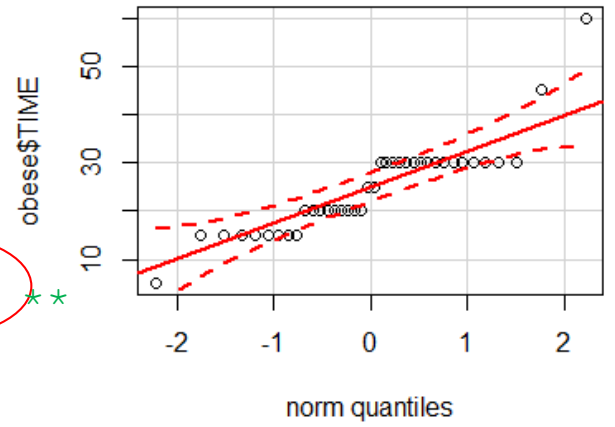


Do medical doctors spend less time with obese patients? How can we test this with linear regression and ANOVA?

```
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

# 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
```



Normality check passed



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

How to test for an effect between >2 groups? Applying 1-way ANOWA with >2 levels

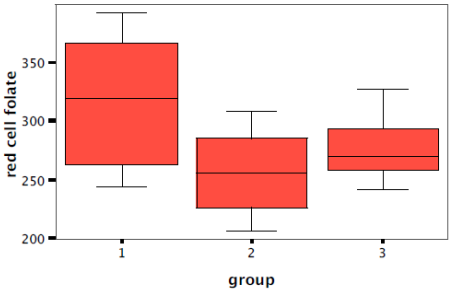
Here, we want to investigate, if three different treatments result in different levels of the output: folate in red blood cells

We can apply a regression with the group factor as predictor to investigate this question, given the folate values y in each group are i.i.d. normal distributed (check not shown).

```
fit=lm(folate~group, data=dat)
anova(fit)     # p=0.044
```

Since p<5%, we can conclude that there are differences, i.e. the folate level is not the same in all groups.

group	red cell folate
1	243
1	251
1	275
1	291
1	347
1	354
1	380
1	392
2	206
2	210
2	226
2	249
2	255
2	273
2	285
2	295
2	309
3	241
3	258
3	270
3	293
3	328



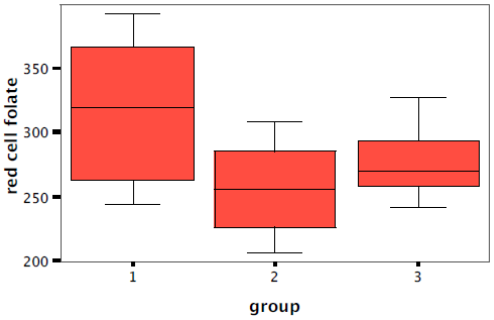
Remark: If there is only 1 factor as predictor, like treatment group, we talk about 1-way ANOVA regardless of the number of groups.

The ANOVA gets significant
Between which groups are the differences?

The significant ANOVA result, only tells us, that there are any differences.
We need to perform post-hoc tests to investigate, between which groups we can really find differences.

We can perform three pair-wise t-tests.
Only the t-test comparing group 1 versus 2 gets significant.

We need to correct for multiple testing, e.g. by Bonferroni-correction. Here, this correction leads to non-significance for all 3 tests.



Result of (uncorrected) pair-wise t-tests:

	Mean Diff.	DF	t-Value	P-Value
1 vs. 2	60.181	15	2.558	0.0218
1 vs. 3	38.625	11	1.327	0.2115
2 vs. 3	-21.556	12	-1.072	0.3046

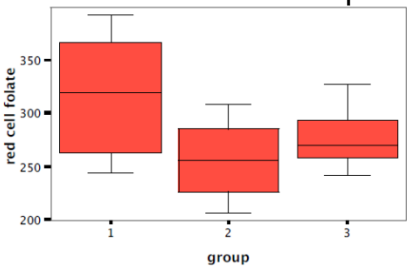
List of post-hoc tests (from wiki)

- Fisher's least significant difference: LSD
- Bonferroni correction
- Duncan's new multiple range test
- Friedman test
- Newman–Keuls method
- Scheffé's method
- Tukey's range test
- Dunnett's test

Non-parametric one-way ANOVA between >2 groups

If outcome-values given a certain predictor-value do not follow a Normal-distribution, we use a non-parametric test.

independent data
All observation are independent



Case1: Data are independent, uncorrelated, un-paired

For the former example, it would look like:

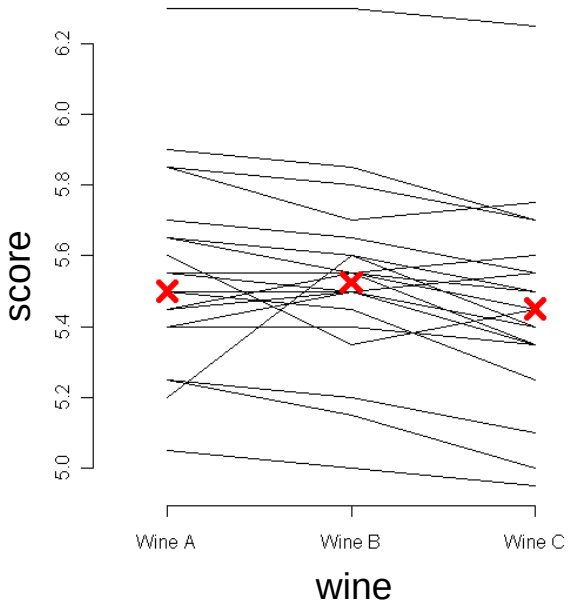
```
kruskal.test(folate~group, data=dat)
```

Case2: Data are dependent, matched, grouped

Three different wines were tasted and scored by 22 people, where each person scored every wine. The data are not independent, since we have a person-grouping. To take account for individual differences in scoring, we perform the friedman-test:

```
friedman.test(Taste ~ Wine | Taster,
                    data=WineTasting)
```

Dependent data
each line correspond to 1 person



Remark: Paired post-hoc tests are needed in addition.

How to assess if there is a change of a numeric output variable when explanatory variables (treatment) change?

Outcome Variable	Parametric tests: The observations are normally distributed under fixed values of the input variables.		Non-parametric tests if the normality assumption is violated or the sample size is small
	un-paired independent	**paired**, dependent, correlated	
Continuous (e.g. pain scale, conc., cognitive function)	**Unpaired t-test=** **1-way ANOVA with 2 groups:** compares means between two independent groups	**Paired t-test:** compares means between two related groups (e.g., the same subjects before and after)	Non-parametric statistics **Wilcoxon sign-rank test**: non-parametric alternative to the **paired** t-test for **2 groups** **Wilcoxon sum-rank test** (=Mann-Whitney U test): non-parametric alternative to the **unpaired** t-test for **2 groups**
	ANOVA: compares means between more than two independent groups: is there any difference between groups? **Pearson's correlation coefficient** (linear correlation): shows linear correlation between two continuous variables **Linear regression:** multivariate regression technique used when the outcome is continuous; gives slopes	**Repeated-measures ANOVA:** compares changes over time in the means of ≥ 2 groups (repeated measurements) **Mixed models/GEE modeling**: multivariate regression techniques to compare changes over time between two or more groups; gives rate of change over time	**Kruskal-Wallis test:** non-parametric alternative to ANOVA for **>2 independent groups**. **Friedman test:** non-parametric alternative to ANOVA **>2 dependent groups**. **Spearman rank correlation coefficient:** non-parametric alternative to Pearson's correlation coefficient

Steps in linear modelling

0) Preprocessing

- learning the meaning of all variables, check for correlations
- give short and informative names
- check for impossible values, errors
- if they exist (missing, error): set them to NA
- be very careful with imputation methods, are missings systematic?

1) First-aid transformations

- bring all variables to a suitable scale (use also field knowledge)
- routinely apply the first-aid transformations

2) Find a good model

- start with a model including important confounders
- perform a residual analysis
- improve model by transformations or adding better predictors
- reduce step by step complexity and use anova for comparison
- use your specific knowledge to choose between variables

Limits of linear Regression

If your residuals do not follow a Normal distribution (even after transformations) use generalized linear modeling
(glm – e.g. logisitic regression)

If your predictors show a strong correlation use shrinkage methods (e.g. lasso)

If your data are not independent use mixed models or methods for time-series.

If you do not have a linear relation, use non-linear regression (e.g. nlm) or generalizes additive models (e.g. gam) or tree models