Biostatistics week 9

➢ Wrap-up simple regression
   - standard errors
   - confidence and prediction interval
   - logged response model

➢ Adjusting for covariables
   - Multiple regression

Setting the scene

Model for the <u>c</u>ondition <u>p</u>robability <u>d</u>istribution

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

$Y_x \in \mathbb{R}$, $\mu_x \in \mathbb{R}$

The predicted value of the linear regression gives only one of the parameter of the CPD: μ_x that depend on the predictor values. The second parameter of the CPD (σ^2) is assumed to be independent of the predictor values and defines the variance of the error term ε.

$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \varepsilon_i$

$E(Y_{X_i}) = \mu_{x_i} = (\mu|X=x_i) = \beta_0 + \beta_1 \cdot x_{i1}$

$Var(Y_{X_i}) = Var(Y|X_i) = Var(\varepsilon_i) = \sigma^2$

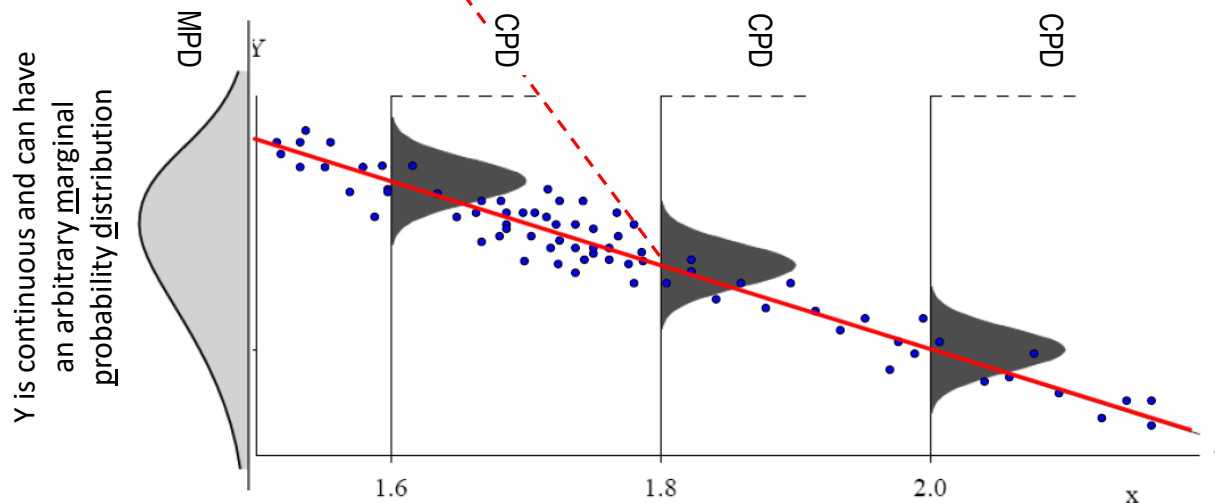ε_i i.i.d. $\sim N(0, \sigma^2)$

<u>i</u>dentical <u>i</u>ndependent <u>d</u>istributed

$Y \sim V_{arbirary}^{contiuous}$

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

MPD

Y is continuous and can have an arbitrary <u>m</u>arginal <u>p</u>robability <u>d</u>istribution

CPD   CPD   CPD

1.6   1.8   2.0   x

2

Use linear regression to predict bodyfat from BMI

```
> r.bodyfat <- lm(bodyfat ~ bmi,d.bodyfat)
> summary(r.bodyfat)

Call:
lm(formula = bodyfat ~ bmi, data = d.bodyfat)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5485  -3.5583   0.0785   4.0384  12.7330

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -26.9844     2.7689  -9.746   <2e-16 ***
bmi           1.8188     0.1083  16.788   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.573 on 241 degrees of freedom
Multiple R-squared:  0.539,        Adjusted R-squared:  0.5371
F-statistic: 281.8 on 1 and 241 DF,  p-value: < 2.2e-16
```
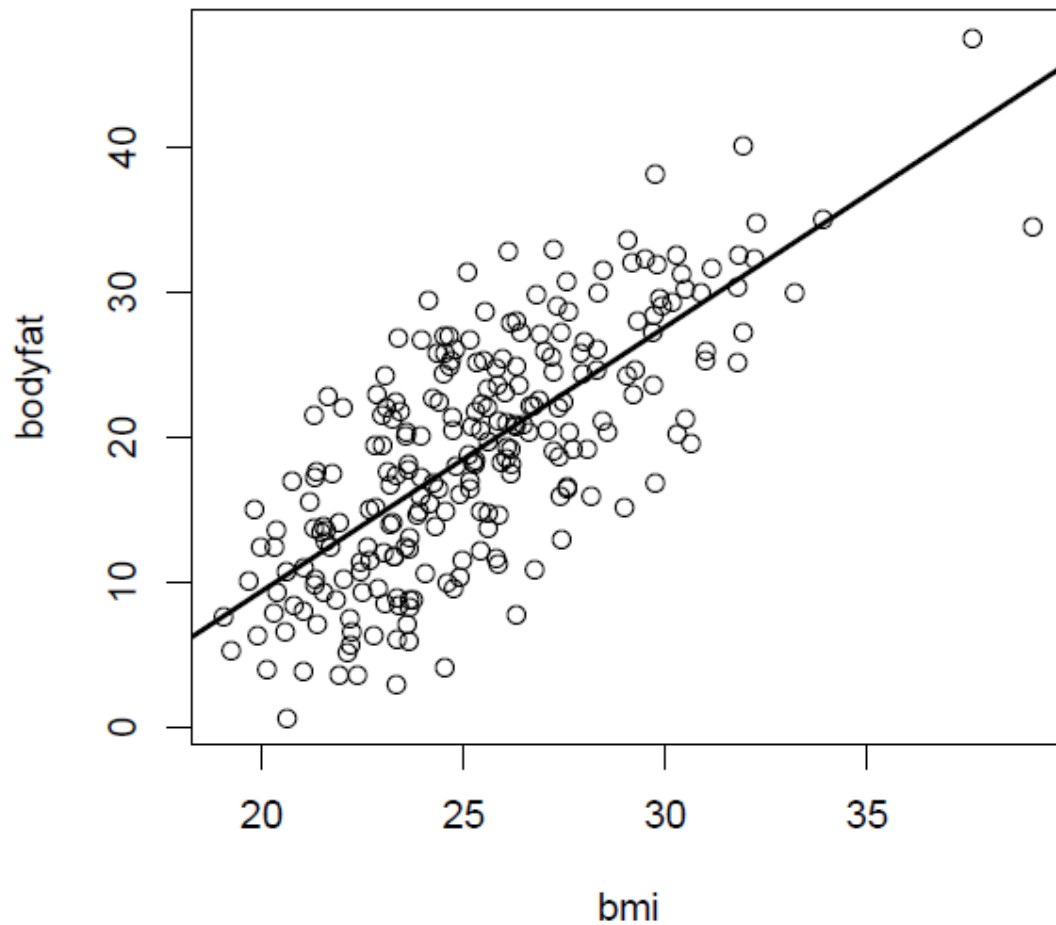
$$\Rightarrow \hat{\alpha} = -26.98 \ , \ \hat{\beta} = 1.82, \ \hat{\sigma}_e = 5.57$$

Plot the regression line into the scatterplot

```
> plot(bodyfat ~ bmi,d.bodyfat)
> abline(r.bodyfat,lwd=2)
```

Assumptions of a linear regression model

Before we continue to look into the results, we need to check if the modelling assumptions are met!

Why? Because otherwise we draw invalid conclusions from the results.

The assumption we took here is that the errors $\varepsilon_i \sim N(0,\sigma^2)$

This implies four things for the residuals which we use as estimates for the errors:

a) The expected value of r_i is 0: $E(r_i)= 0$.

b) All r_i have the same variance: $Var(r_i) = \hat{\sigma}^2$.

c) The r_i are normally distributed.

d) The r_i are independent of each other.

Model checking

The Tukey-Anscombe diagram plots the residuals against the fitted values (see left plot) is the most important model checking tool. This plot is ideal to check if assumptions a) and b) (and partially d)) are met. Normality of the residuals is assumption c) and can be easily checked – e.g. by inspecting the Normal-Q-Q plot of the residuals (see right plot). Here, all is o.k..
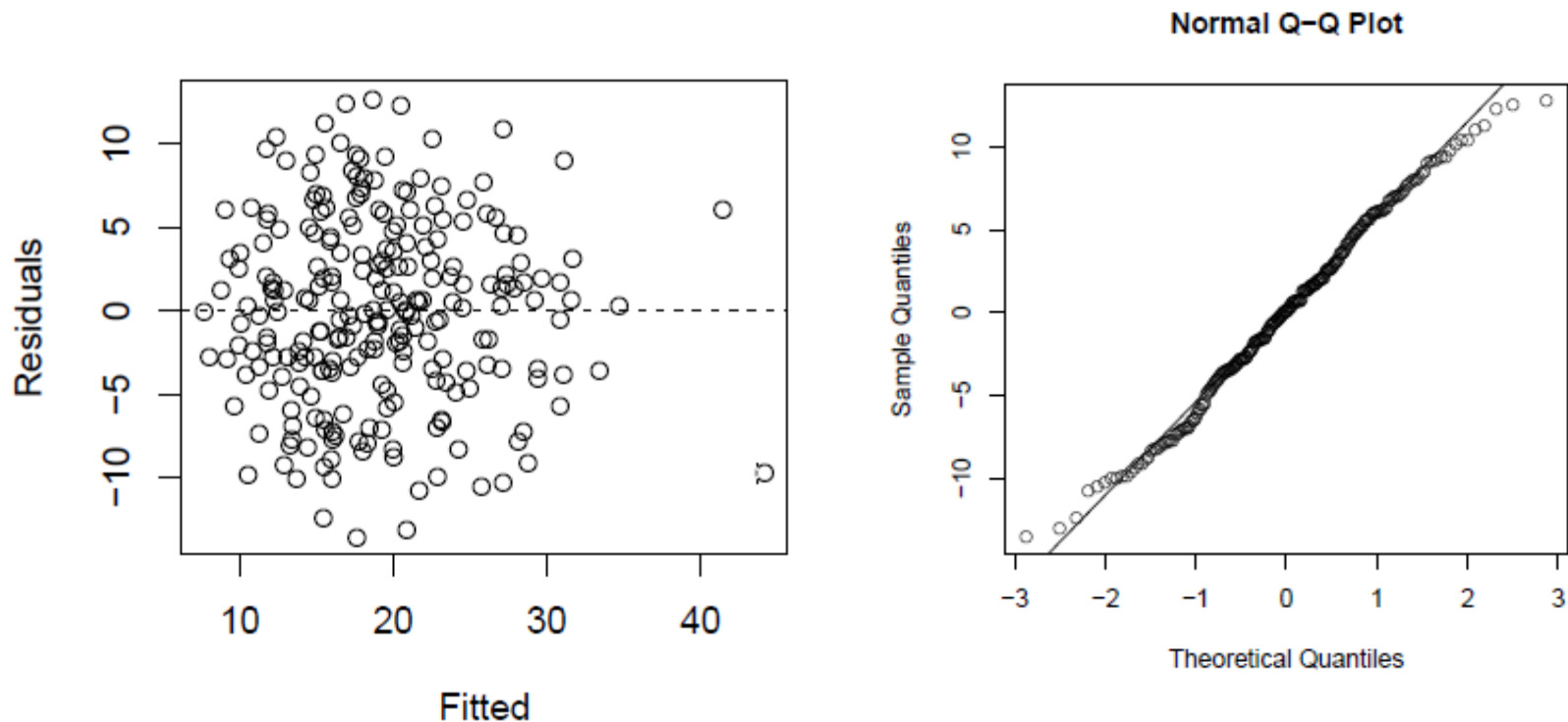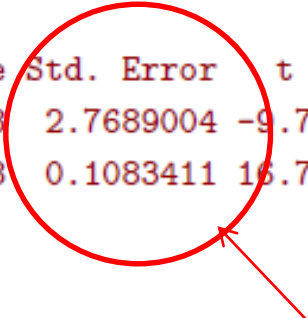
Uncertainty of the estimated model coefficients

Lets look at the estimates and the standard errors of the model coefficients:

```
> summary(r.bodyfat)$coef

             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) -26.984368  2.7689004 -9.745518 3.921511e-19
bmi           1.818778  0.1083411 16.787522 2.063854e-42
```

The second column shows a standard error of the estimated model coefficient quantify the uncertainty of the estimated coefficients.

If we fit the model with a new sample of data drawn from the same population, we expect to get slightly different estimates.
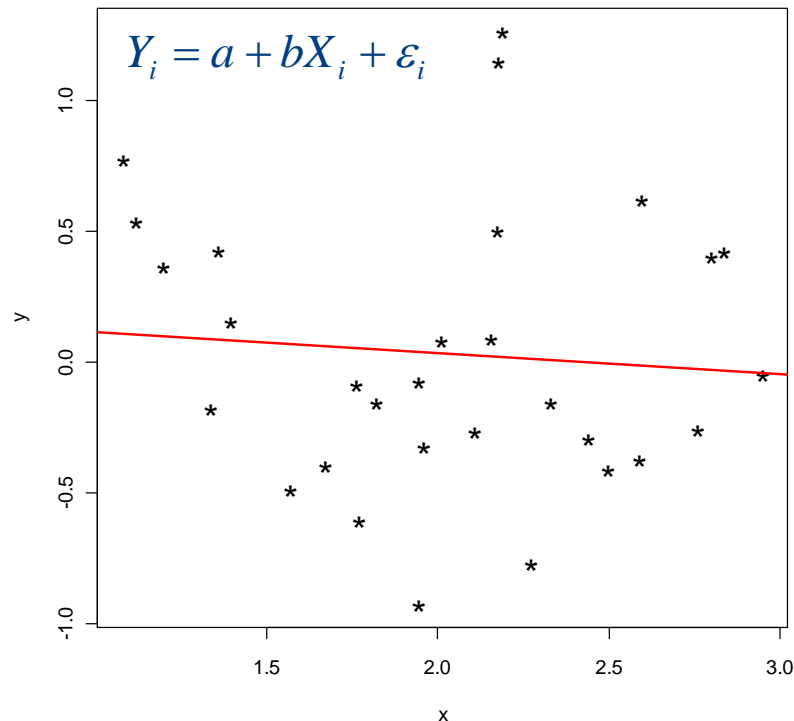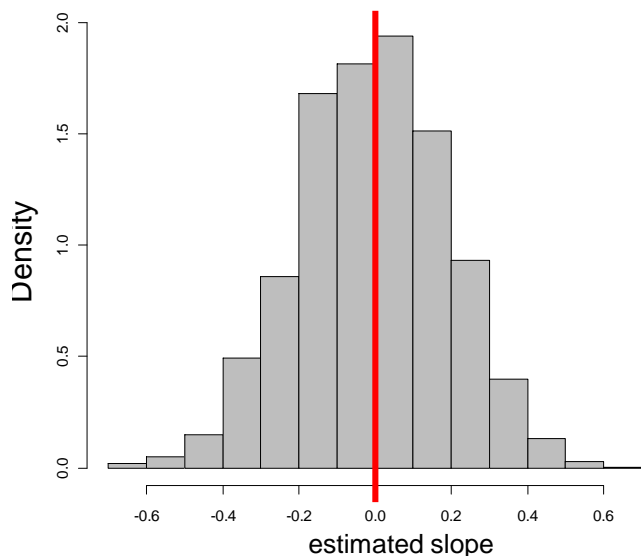
How does the distribution of the estimated coefficients look like?

Determine the distribution of the coefficients under H_0

```
x=runif(n=30,min=1,max=3  )
my.err = rnorm(n=30,mean=0,sd=0.5)
y=0*x+0+my.err
plot(x,y,pch="*",cex=2)
mod.l=lm(y~x)
abline(mod.l,lwd=2,col="red")
coef(mod.l)[2]  # -0.0049
```

Repeat simulation under H_0 5000 times:

$$\hat{b} \sim N\left(0, se(\hat{b})^2\right)$$

$$Y_i = a + bX_i + \varepsilon_i$$

The estimated coefficients are normally distributed with expected value equal to the true coefficients which have been used to simulate the data.

8

Standard errors and confidence intervals of the coeffients

Intercept

$$\hat{a} \pm q^{t_{n-2}}_{\alpha/2} \cdot se(\hat{a})$$

$$se(\hat{a}) = \sqrt{\hat{\sigma}_\varepsilon^2 \cdot \left(\frac{1}{n} + \frac{\overline{x}}{\sum (x_i - \overline{x})^2}\right)}$$

Slope

$$\hat{b} \pm q^{t_{n-2}}_{\alpha/2} \cdot se(\hat{b})$$

$$se(\hat{b}) = \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

<u>remark</u>: the estimates are normally distributed- still we have to use the quantiles of the t-distribution for the CI since we need to estimate the variance of the normal distribution - the formulas for the se are not relevant for the exam.

Testing the Slope

There is a statistical hypothesis test which can be used to check whether the slope is significantly different from zero:
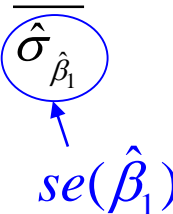
b

$$H_0 : \beta_1 = 0$$

One usually tests two-sided on the 95%-level. The alternative is: $H_A : \beta_1 \neq 0$

As a test statistic, we use:

$$T_{H_0:\beta_1=0} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \qquad T \sim t_{n-2}$$

$$se(\hat{\beta}_1)$$

The summary of lm in R gives the t-value and the respective p-value of this test.

The summary provides all information needed to do manually a test for H_0: $\beta = b$

$$T_{H_0:\beta_1=b} = \frac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}}$$
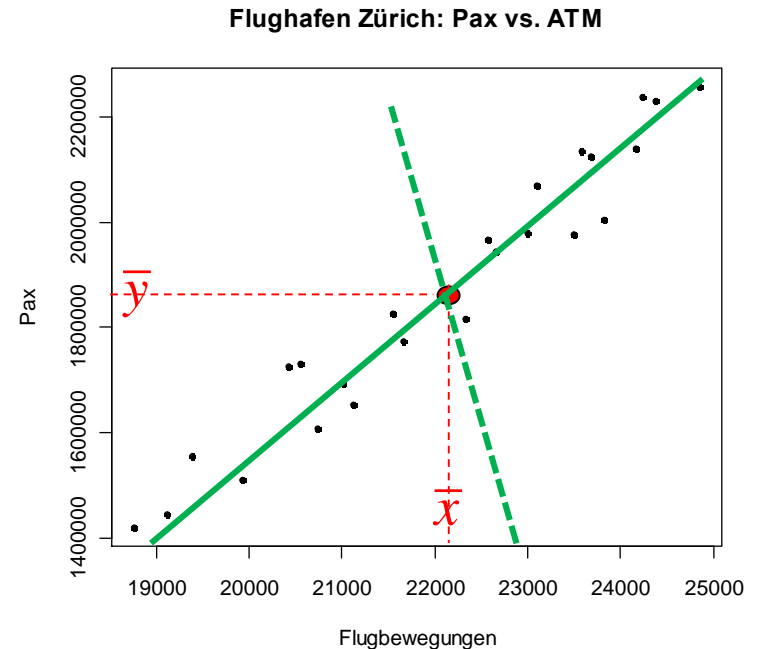
Testing the Intercept

An analogous test can be done for the intercept.

➢ No matter what the test result will be, the intercept should generally not be omitted from the regression model.

➢ The presence of an intercept protects against possible non-linearities and calibration errors of measurement devices. If it is kicked out of the model, the results are generally worse.

➢ If theory dictates that there should **not** be an intercept but it is still significant, take this as evidence that the linear relation does not hold when extrapolating to $x=0$.

The linear model goes through the center of mass

The sum of the residuals need to be zero – otherwise the model is biased.
But this is not suffient, because also quite bad models can have this property.

$$0 = \sum_{i=1}^{n} r_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))$$

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \alpha - \beta \sum_{i=1}^{n} x_i$$

$$= n\bar{y} + n\alpha - n\beta\bar{x}$$

$$\Longleftrightarrow \quad \bar{y} = \alpha + \beta\bar{x}$$

Flughafen Zürich: Pax vs. ATM

Standard errors and confidence intervals of the outcome

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \qquad \hat{\sigma}_\varepsilon = \frac{1}{n-2} \sum_{i=1}^{n} r^2_i$$

Estimated outcome and its CI:

$$se(\hat{\sigma}_\varepsilon) = \sqrt{\hat{\sigma}_\varepsilon^2 \cdot (1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2})}$$

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} \cdot x_i$$

$$(\hat{a} + \hat{b} \cdot X_k) \pm q^{t_{n-2}}_{\alpha/2} \cdot se(\hat{y}) \qquad se(\hat{y}(x)) = \sqrt{\hat{\sigma}_\varepsilon^2 \cdot (\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2})}$$

Width of CI depends on x and is smallest at x=mean(x)

remark: the estimates are normally distributed- still we have to use the quantiles of the t-distribution for the CI since we need to estimate the variance of the normal distribution - the formulas for the se are not relevant for the exam.

Confidence- and prediction intervals

Remember: We estimate the regression coefficients from a random sample. There are two questions to be asked:

- How to give an interval so that the risk that the true linear model runs not within this interval is not larger than 5%?

  →This leads to the confidence range.

- How to give an interval so that the risk that an individual observation is not within interval is not larger than 5%?

  → This leads to the prediction range.
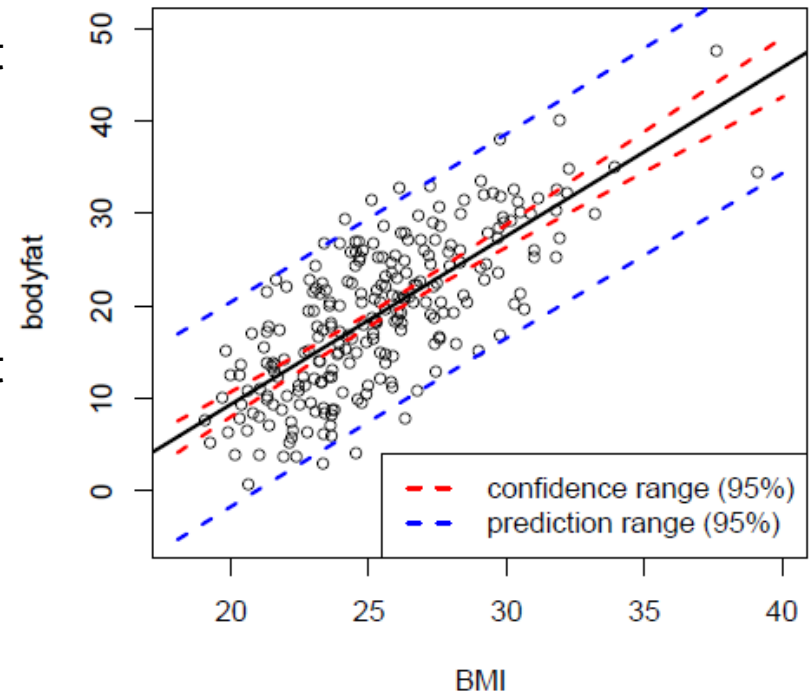
Confidence- and prediction intervals
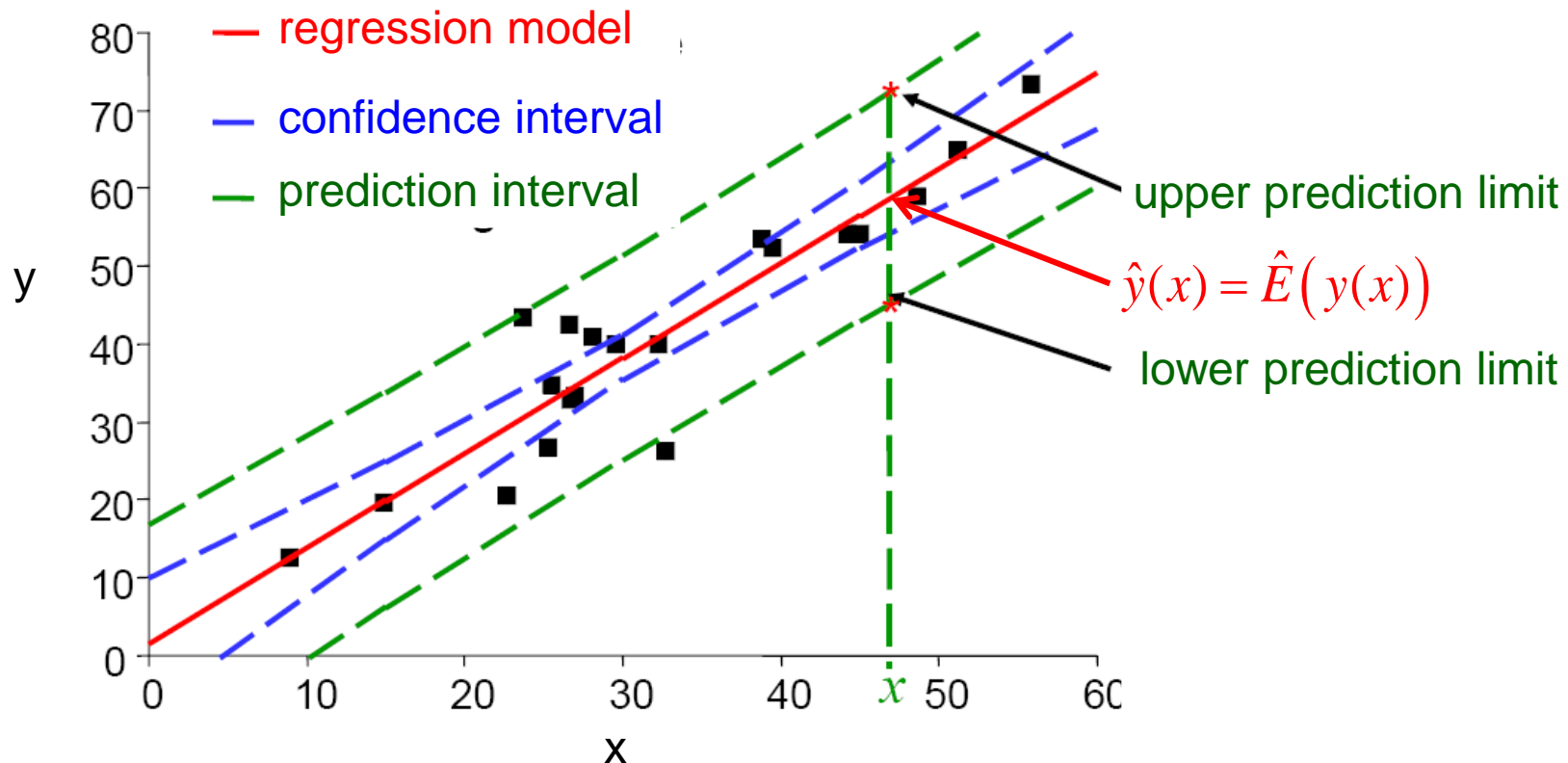
Regression yields an estimate for the expected value of y for each x value and the true <u>expected value</u> of y at the position x is with 95% percentage certainty covered by the confidence interval.
A <u>single observation</u> y at the position x is with 95% percentage certainty within the prediction interval.

Interesting intervals when doing regression

$$Y_i = a + bX_i + \varepsilon_i$$

CI for y-intercept:	$\hat{a} \pm q^{t_{n-2}}_{\alpha/2} \cdot se(\hat{a})$	➢confint(fit, parm=1,level=0.95) 2.5 % 97.5 % (Intercept) -32.438703 -21.530032
CI for the slope:	$\hat{b} \pm q^{t_{n-2}}_{\alpha/2} \cdot se(\hat{b})$	➢confint(fit,parm=2,level=0.95) 2.5 % 97.5 % x 1.605362 2.032195
CI for $E(y_k)$:	$\hat{y}_k \pm q^{t_{n-2}}_{\alpha/2} \cdot se(\hat{y}_k)$	➢predict(fit, new, se.fit=T, interval=c(**"confidence"**)) fit lwr upr 23 22.70 23.3
Prediction interval for yl:	$\hat{y}_k \pm q^{t_{n-2}}_{\alpha/2} \cdot \left(\hat{\sigma}_\varepsilon + se(y_k)\right)$	➢predict(fit, new, se.fit=T, interval=c(**"prediction"**)) fit lwr upr 23 21.20 24.80

Plotting Confidence and Prediction Interval

Note:

Visualizing the confidence and prediction intervals in R is
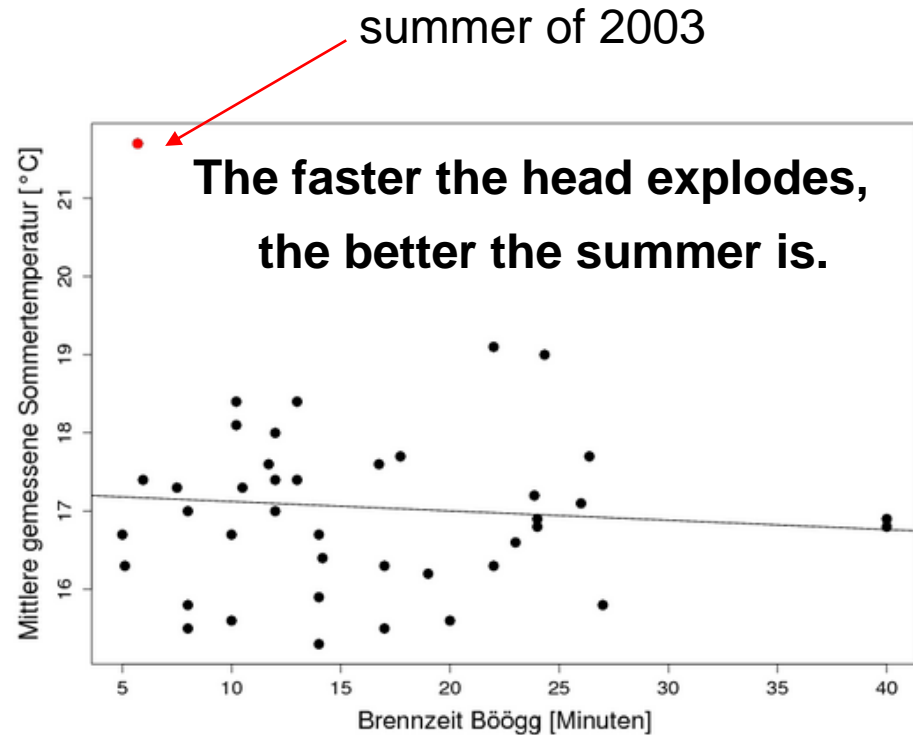*not straightforward, but requires some tedious handwork.*

New data for prediction must be provided in a data frame
and the columns with the predictors must have the same
names as in the data frame that has been used in fit.

R-Hints:

```
dat  <- data.frame(x=seq(...,..., length=200))
pred <- predict(fit, newdata=dat, interval=...)
plot(..., ..., main="...")
lines(dat$x, pred[,2], col=...)
lines(dat$x, pred[,3], col=...)
```

The „Böögg" as weather oracle

summer of 2003

**The faster the head explodes,
the better the summer is.**

outliers violate the model assumption of i.i.d. normal distributed residuals! i.e. linear regression as we have seen until now is not a suitable paradigm anymore.

Outlier are dangerous!
## especially when they occur at the border of the x-range
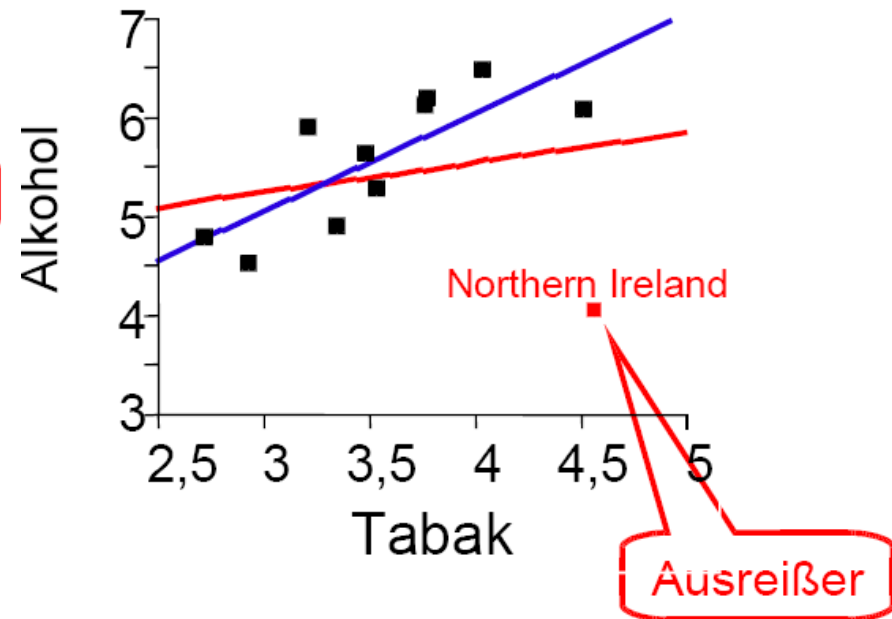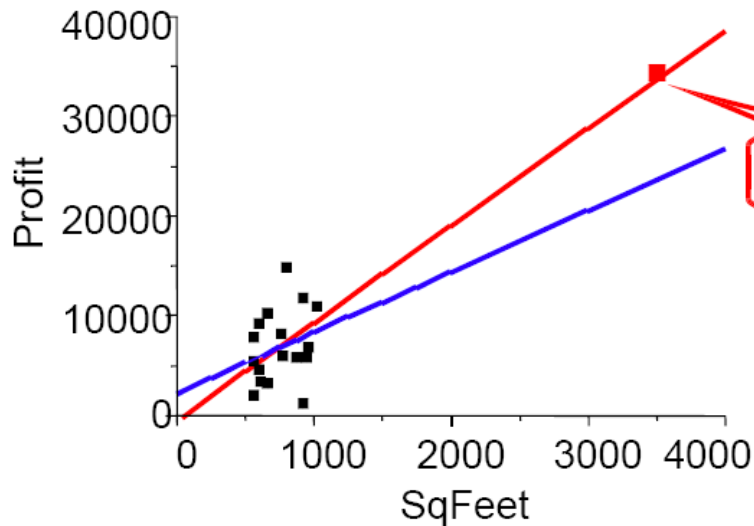
An Outlier can have giant impact on the result, especially when it occurs at the border of the x-range (leverage point). The result can get totally influenced by such an observation and therefore be useless. In such cases one can apply roubst regression (rlm(…)).

In which cases should the dependency between two variables be described by linear regression
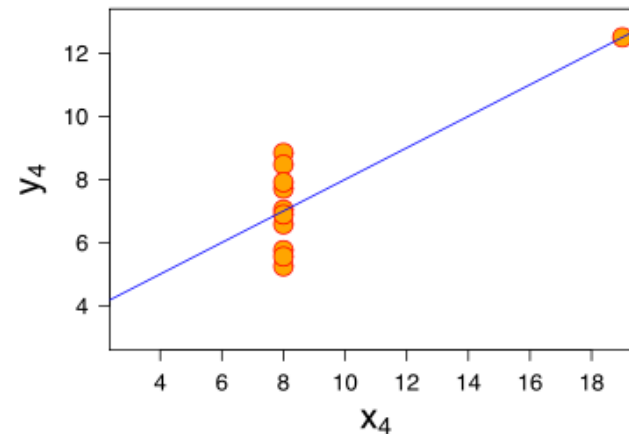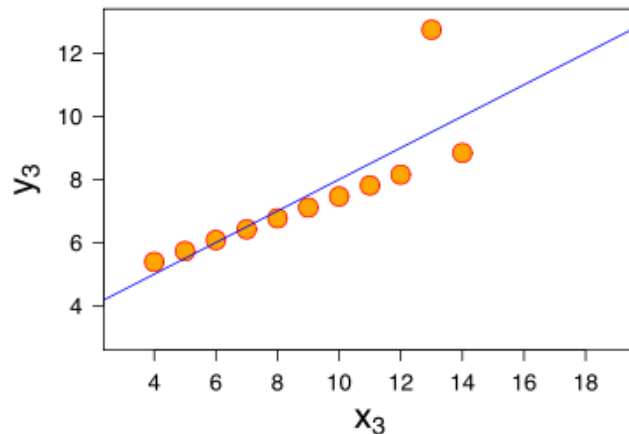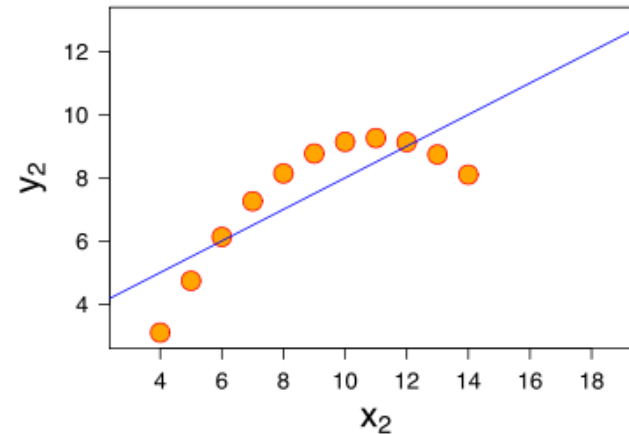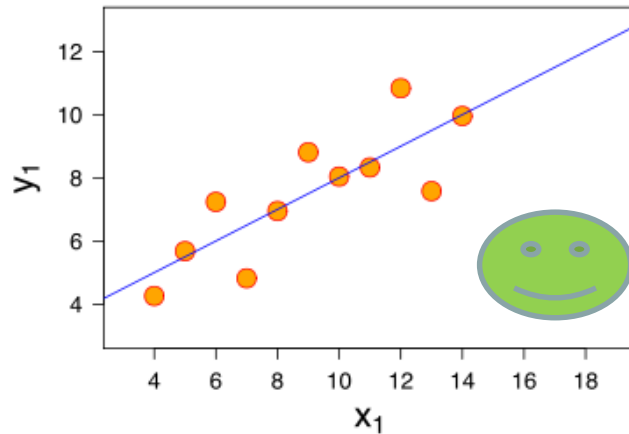
All depicted datasets lead to the same regression fit…
Never fit a regression fit without check the model diagnostics.

Examples of Tukey-Anscombe-Plots
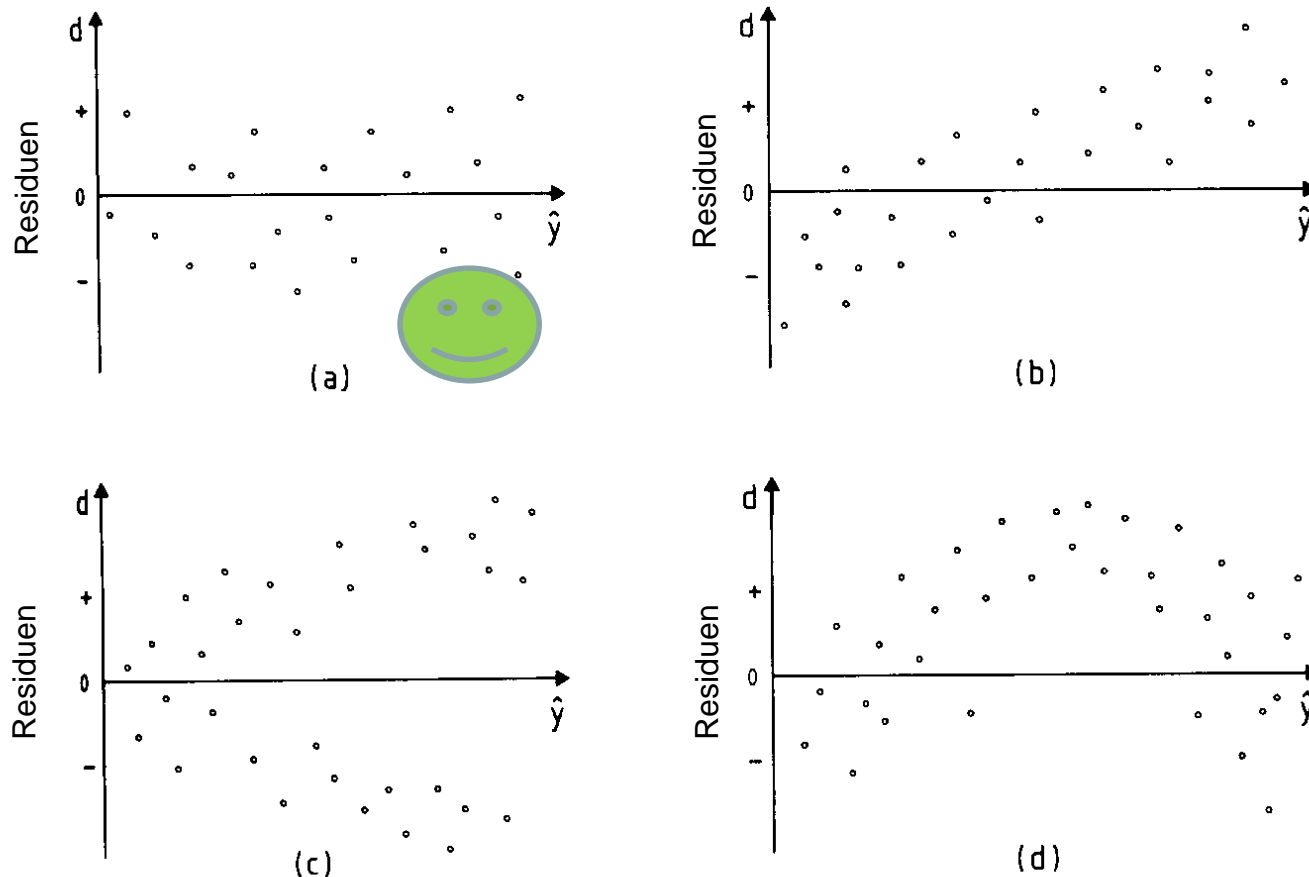
Abb. 12: Graphische Darstellung der normierten Residuen in Abhängigkeit von den geschätzten Werten ŷ des Regressanden: (a) idealer Verlauf; (b) linearer Trend, auf einen Rechenfehler hindeutend; (c) ansteigende Varianzen (mit ŷ), Wechsel zu einem Modell für ungleiche Varianz der Beobachtungen eventuell angebracht; (d) nichtlinearer Verlauf der Residuen, inadäquates Modell, d. h. Transformation der Daten oder Änderung der Regressionsfunktion angezeigt

Additive and multiplicative error terms

model with additive error term (assumption for linear regression):

$$Y_i = \underbrace{a + b \cdot X_i}_{} + \varepsilon_i$$

systematic
part of the model

random part
of the model

A model with a multiplicative error term.

$$Y_i = \underbrace{f(x)}_{} \cdot \varepsilon_i$$

systematic
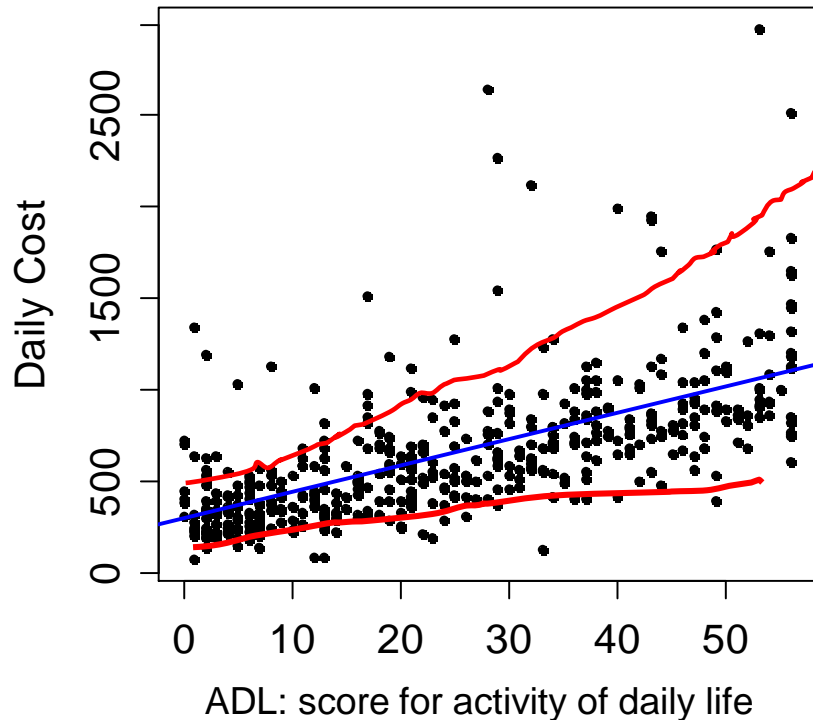part of the model

random part
of the model

can be rewritten as a model with additive errors by using the logarithm.
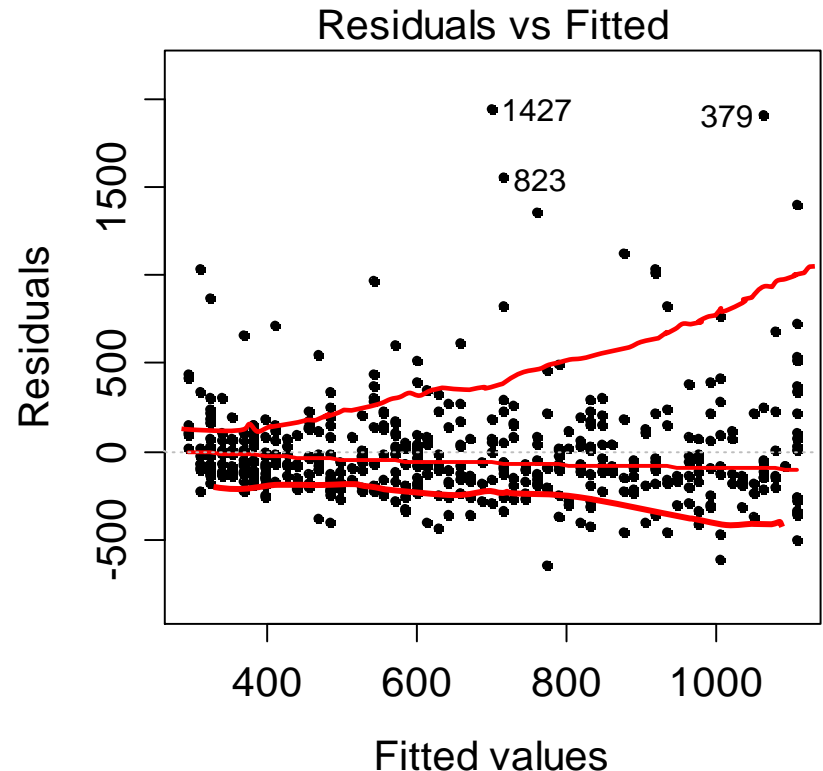
$$\log(Y_i) = \log(f(X)) + \log(\varepsilon_i)$$

random part
of the model

What can we do, if residual variance is not constant?

Daily Cost in Rehab vs. ADL

Daily Cost vs ADL: score for activity of daily life

Residuals vs. Fitted Values

Residuals vs Fitted

1427   379
823

Residuals vs Fitted values

In case of multiplicative errors we see usually a funnel shape in the data/residuals. We already see in the original data that the variance is not constant and the points lay within a funnel. A funnel is visible in the residual plot together with some extreme residuals. Hence, the assumption for the linear regression are violated.

Logged Response Model

We transform the response variable and try to explain it using a linear model with our previous predictors:

$$Y' = \log(Y) = \beta_0 + \beta_1 x + E$$

In the original scale, we can write the logged response model using the same predictors:

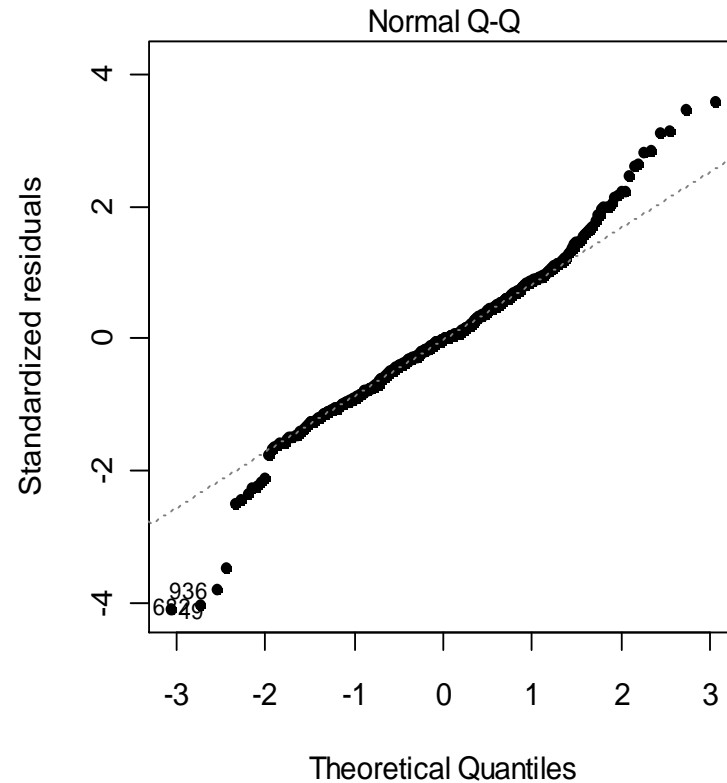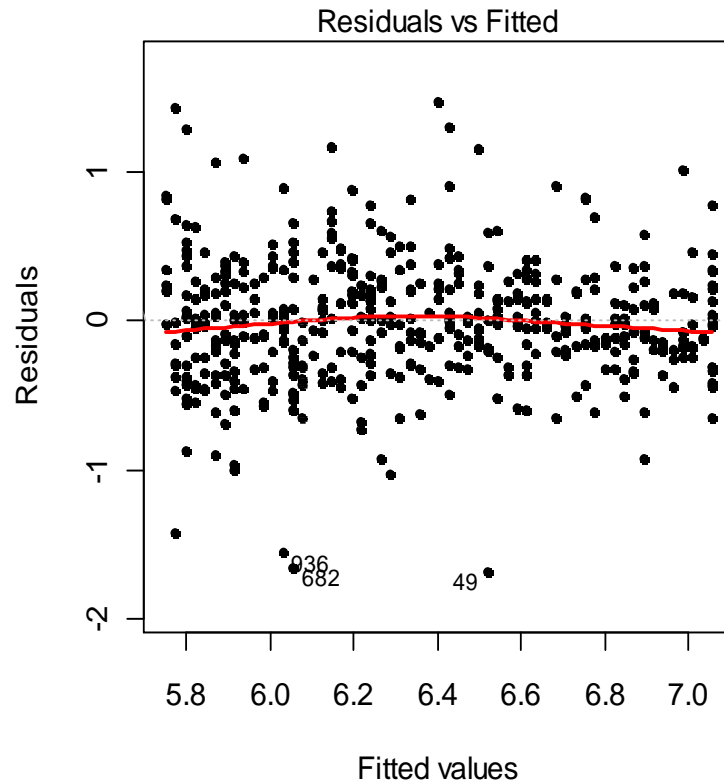$$Y = \exp(\beta_0) \cdot \exp(\beta_1 x) \cdot \exp(E)$$

→ Multiplicative model

→ $E \sim N(0, \sigma_E^2)$, and thus, $\exp(E)$ has a lognormal distribution

Log-Transformation has stabilized the variance!

$$Y' = \log(Y) = \beta_0 + \beta_1 x + E$$

Back Transforming the Fitted Values

- In principle, we can „simply back transform" – under this transformation we get from a straight line to an exponential curve

$$\hat{y} = \exp(\hat{y}')$$

- Since the exp-transformation is a monotonic function, the quantile transform also with exp and the limits of confidence/prediction intervals are determined accordingly

$$[l, u] \;\rightarrow\; [\exp(l), \exp(u)]$$

- $\widehat{y'}$ is an estimate for the mean (=median because of symmetric distribution) of the loged outcome log(y), however (since only quantiles as the median, but not the mean, transform under a monotonic non-linear function as exp-function) $\hat{y}$ is an estimate for the median (not the mean) of the outcome y (here: cost) conditioned on the values of the input variables (here: ADL)

Model and Prediction Intervall
after Back-Transfomation

Daily Cost in Rehabilitation vs. ADL-Score
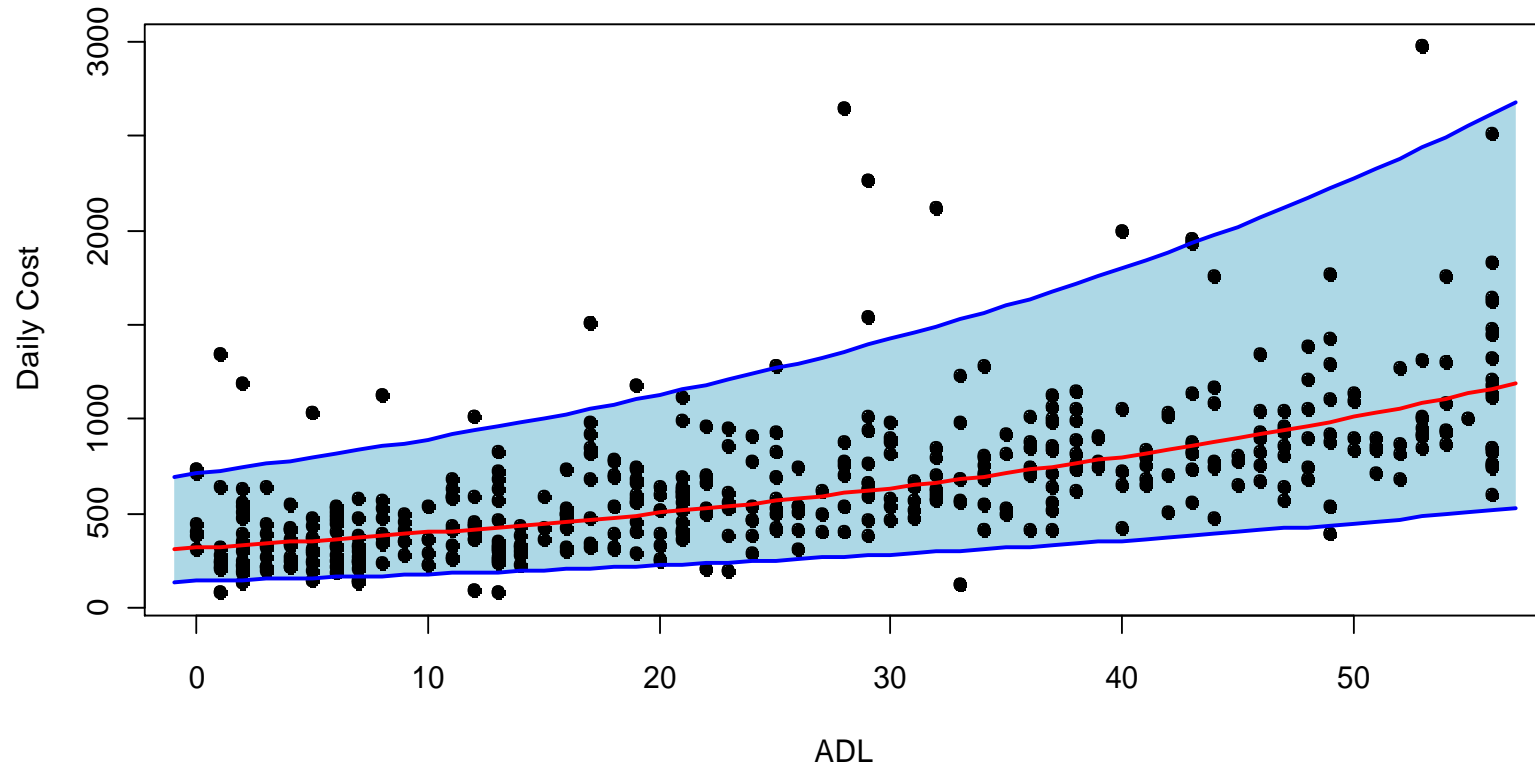
Interpretation of the Coefficients in a logged Response Model

Important: there is no back transformation for the coefficients to the original scale, but still a good interpretation

$$\hat{y}' = \log(\hat{y}) \quad = \quad \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$$

$$\hat{y} \quad\quad\quad = \quad \exp(\hat{\beta}_0) \cdot \exp(\hat{\beta}_1 x_1) \cdot ... \cdot \exp(\hat{\beta}_p x_p)$$

An increase by one unit in x_i would add $\widehat{\beta}_i$ to $\widehat{y}' = \log(\hat{y})$, but it would multiply the fitted value $\hat{y}$ in the original scale with $e^{\widehat{\beta}_i}$.

→ **Coefficients are interpreted multiplicatively on original scale! If the input variable x_i increases by 1 unit (all other input variables are fixed), that the output variably $\hat{y}$ changes by the factor $e^{\widehat{\beta}_i}$.**

First Aid Transformations
Variance-stabilizing transformations

These transformations intended to stabilize the variance

First-Aid Transformations:

→ do always apply these (if no practical reasons against it)

→ to both response and predictors

Absolute values and concentrations:

log-transformation: $y' = \log(y)$

Count data:

square-root transformation: $y' = \sqrt{y}$

Proportions:

arcsine transformation: $y' = arc\sin\left(\sqrt{y}\right)$

How can we adjust for other influences?

| From Table 3	Entire sample (n=140)	Group with ≥85% consumption (n=130)
Model 1 Unadjusted		
Intercept	−0·31 (−0·49 to −0·13)*	−0·33 (−0·53 to −0·13)†
Challenge type		
Mix A vs placebo	0·20 (0·01 to 0·40)‡	0·24 (0·02 to 0·47)‡
Mix B vs placebo	0·16 (−0·04 to 0·35)	0·16 (−0·07 to 0·38)
Model 2 Adjusted		
Intercept	−0·54 (−0·89 to −0·18)*	−0·51 (−0·92 to −0·11)
Challenge type		
Mix A vs placebo	0·20 (0·01 to 0·39)‡	0·28 (0·05 to 0·51)‡
Mix B vs placebo	0·17 (−0·03 to 0·36)	0·19 (−0·04 to 0·41)

What does "adjusted" mean?

How is it done?

In model 2, in addition to challenge type, the effects of the following factors were adjusted for: week during study, sex, GHA in baseline week, number of additives in pretrial diet, maternal educational level, and social class.

30

How to adjust for possible confounders if working with observational data?

- ➢ Multiple regression model approaches

- ➢ Stratification / Matching (study desing)

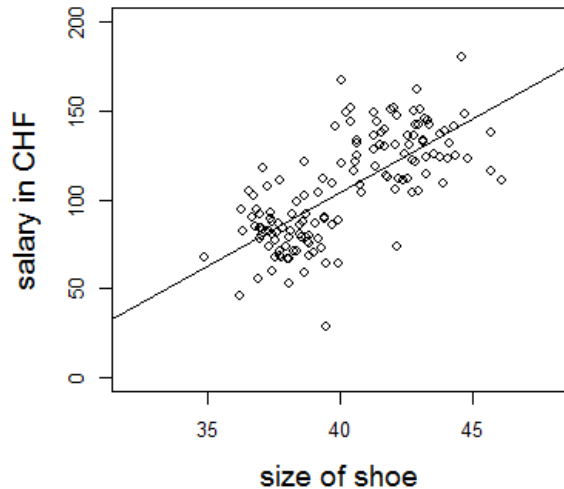- ➢ Propensity Score Adjustment (not covered in this lecture)

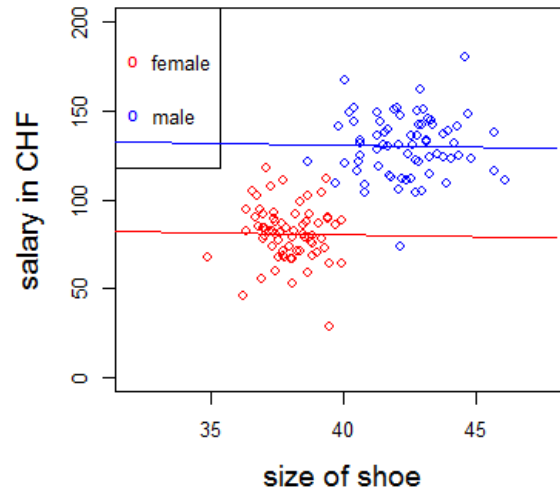# Stratification (or Matching) is most often used in case of few confounders with few levels
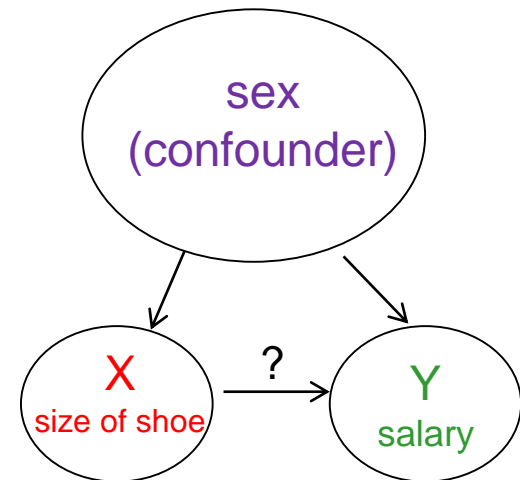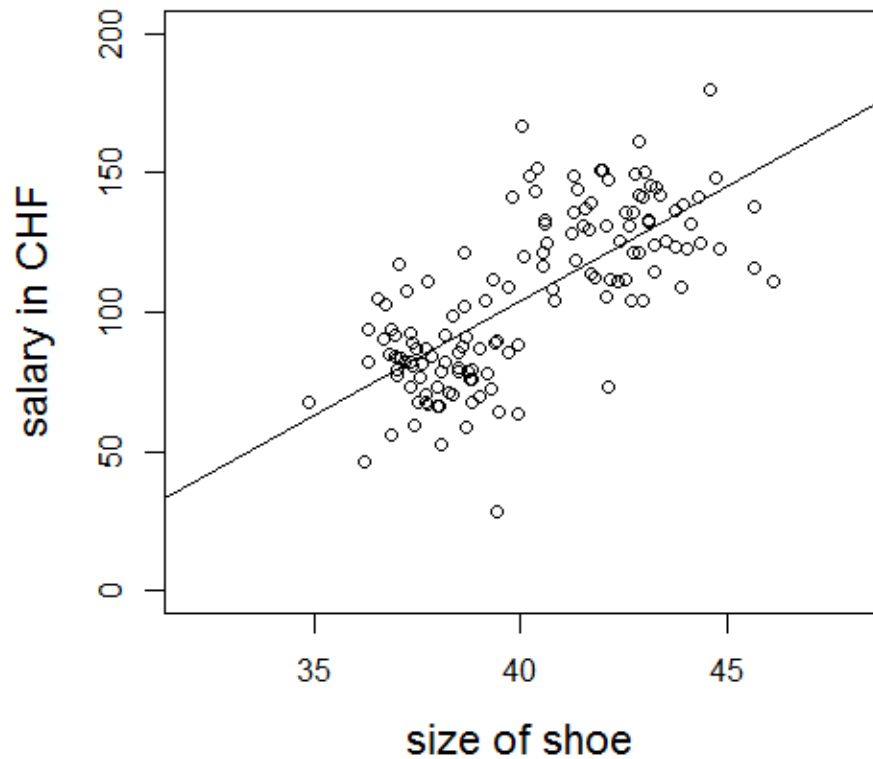
Association between salary and shoe size
before and after taking the sex factor into account

unadjusted on full data set: lm(salary~shoe)

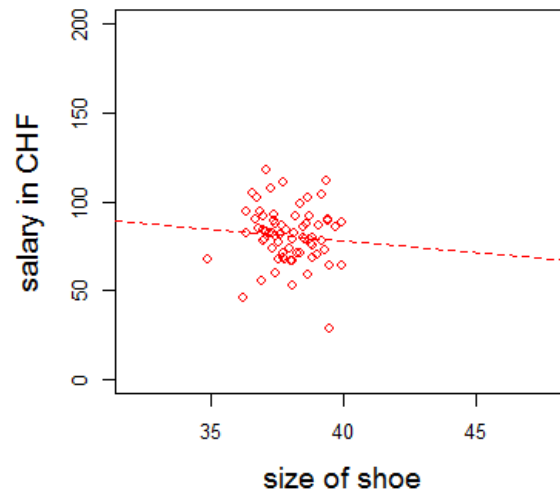adjusted on full data set: lm(salary~shoe+sex)

Adjusting via multiple regression without interaction leads only similar results as a stratified analysis if the interaction is not significant -> parallel regression lines are «feature of model»

males sratum: lm(salary~shoe)

females stratum lm(salary~shoe)

Stratified analysis -> different models for male and females are possible, but here not necessary.

Multiple linear regression: interpretation of coefficient

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \ldots + \beta_p \cdot x_{ip} + \varepsilon_i \quad , \quad \varepsilon_i \sim N\left(0, \sigma^2\right)$$

Matrix Notation:

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{with } \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \ \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The coefficient β_k gives the change of the outcome y, given the explanatory variable x_k is increased by one unit and all other variables are held constant.

$$\beta_k = y_{x_k+1} - y_{x_k} = \Delta y_{x_k -> x_{k+1}}$$

Regression with tow predictor variables

LHCY is the Y (homocysteine) to be predicted from the two X's: LCLC (folate) and LB12 (B_{12}).
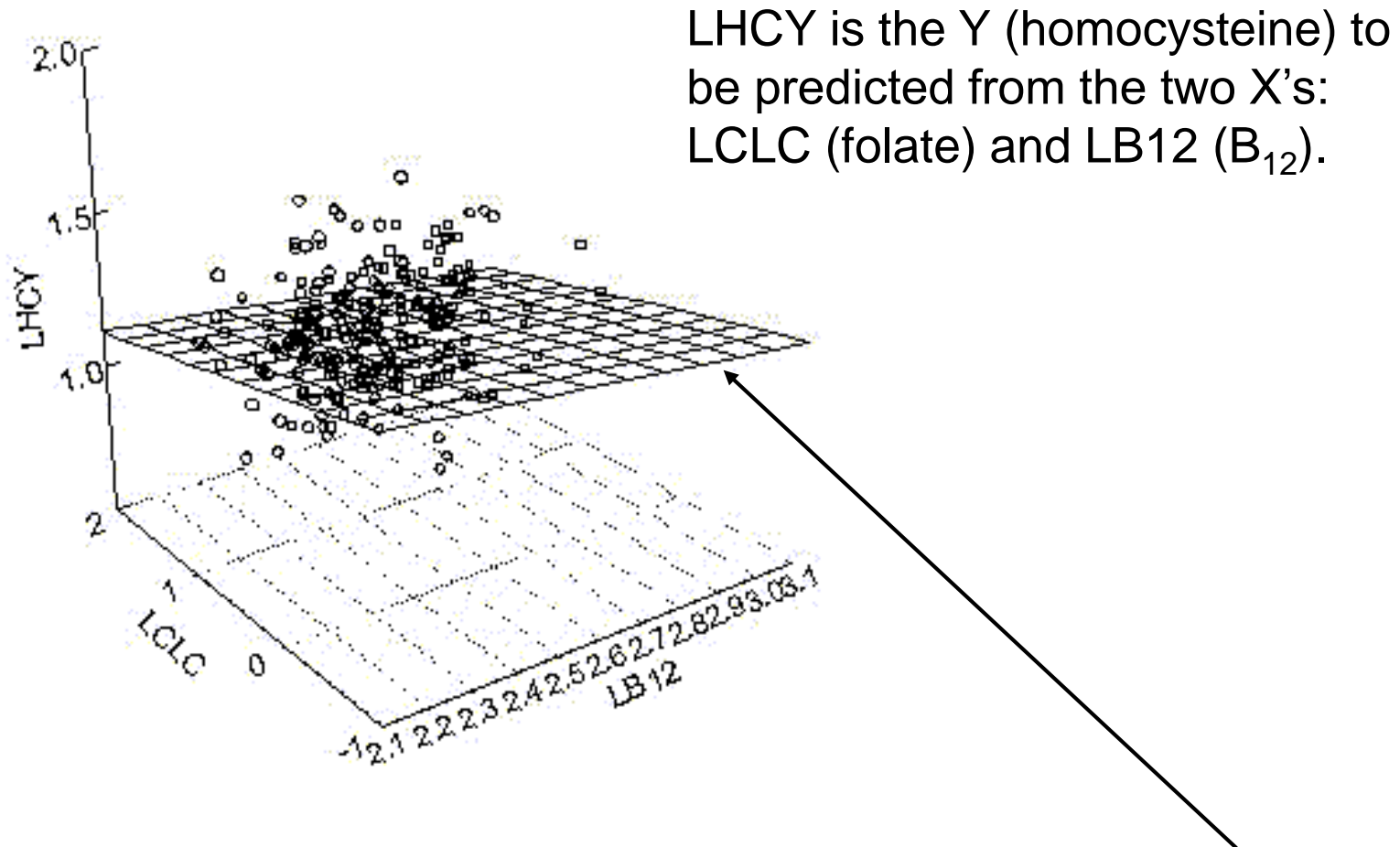
$LHCY = b_0 + b_1 LCLC + b_2 LB12$ is the equation of the plane

Remark: If we have more than 2 predictors the model is a hyper-plane

Example for multiple Regression

Multiple Lineare Regression

LDL（Bad）
HDL（Good）

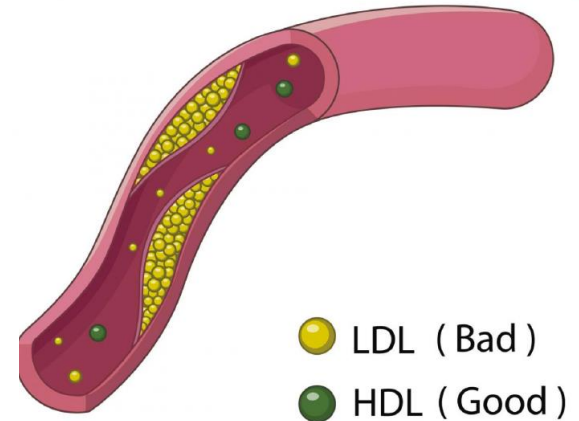Outcome we want to model:

High density lipoprotein (HDL)

Research question:

Which predictors have an impact on HDL?

Preparation:

Make analysis plan

Collect data

Example for multiple regression: HDL example

	Estimate	Std. Error	t-value	Pr(>|t|)
Intercept	1.16448	0.28804	4.04	<.0001
AGE	-0.00092	0.00125	-0.74	0.4602
BMI	-0.01205	0.00295	-4.08	<.0001
BLC	0.05055	0.02215	2.28	0.0239
PRSSY	-0.00041	0.00044	-0.95	0.3436
DIAST	0.00255	0.00103	2.47	0.0147
GLUM	-0.00046	0.00018	-2.50	0.0135
SKINF	0.00147	0.00183	0.81	0.4221
LCHOL	0.31109	0.10936	2.84	0.0051

The predictors of log(HDL) are age, body mass index, blood vitamin C, systolic and diastolic blood pressures, skinfold thickness, and the log of total cholesterol. The equation is:

Log(HDL) = 1.16 - 0.00092(Age) -0.012(BMI)+…+ 0.311(LCHOL)

Linear regression: interpretation of coefficient

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + ... + + \beta_p \cdot x_{ip} + \varepsilon_i$$

The coefficient β_k gives the change of the outcome y, given the explanatory variable x_k is increased by one unit and all other variables are hold constant.

$$\beta_k = y_{x_k+1} - y_{x_k} = y_{x_k -> x_{k+1}}$$

E.g., expected log(LHDL) is 0.012 lower in a subject whose BMI is 1 unit greater, but is the same as the other subject on other factors.

log(HDL) = 1.16 - 0.00092(Age) **-0.012(BMI)**+…+ 0.311(LCHOL)

The meanings of the p-values and the coefficients in the multiple linear regression output

The p-values measure the significance of the association of a factor with Log(HDL) in the presence of all other predictors of the model – meaning "after accounting for other factors" or "adjusting for other factors", and is called independent association.

SKINF alone probably is associated. However, its p=0.42 says that it provides no *additional* information that helps to predict LogHDL, after accounting for other factors such as BMI.

The p-value and also the coefficient-value of a predictor depend i.g. not only on the association with the outcome variable but also on the other predictors in the model.

Only if all predictors are independent multiple regression leads the same p-values and coefficients than p simple regression each with only one predictor.

Significance vs. Relevance

The larger a sample, the smaller the p-values **for the very same predictor effect. Thus** **do not confuse a small p-values with an important predictor effect**!!!

More important than p-values:

- **Look at absolute values of (significant) coefficients.**

- **Look at confidence intervals!**

Summary

- Minimizing squared errors yields an unbiased model describing the train data.

- The 95%-CI of the modeled y gives a range so that we only have a risk of 5% that the true expected value of y is not within this range

- The 95% prediction interval of the modeled y gives a range so that 95% of the fitted data is within this range.

- For a model with p parameters fitted to n observations the 95%CI of each parameter is given by $\hat{b} \pm q^{t_{n-p}}_{\alpha/2} \cdot se(\hat{b})$ which is significantly different from all values not within this CI.

- The larger n the smaller the CI and the more significant parameters we get.

- The parameter size indicates the change of y if the respective variable is increased by one unit and all other variables kept constant (often not possible)

- The size of the parameter and the scale of the variable determine the relevance of a significant parameter

- The size and the p-value of a parameter in a multiple regression model can change if other variables are added to or deleted from the model