Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Lecture 4: Multiple linear regression (finalize) / Residual analysis /
Checking modelling assumptions
16./17. March 2017

Overview

- Interactions between covariates
- Multiple vs. many single regressions
- Checking assumptions / Model validation
- What to do when things go wrong?
- Transformation of variables/the response
- Handling of outliers

Course material covered today

The lecture material of today is based on the following literature:

- Chapters 3.2u-x, 3.3, 4.1-4.5 in *Lineare Regression*

Recap of last week

- Multiple linear regression model
  $y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + e_i.$
- Binary and factor covariates: Introduce dummy variables such that

$$x_i^{(j)} = \begin{cases} 1, & \text{if the } i\text{th observation belongs to group } j. \\ 0, & \text{otherwise.} \end{cases}$$

- Include $x^{(2)}$, ... ,$x^{(k)}$ in the regression, given that $x^{(1)}$ is used as reference category ($\beta_1 = 0$).
- The F-test is used to test if $\beta_2 = \beta_3 = ... = \beta_k = 0$ at the same time for a factor covariate with k levels. Use the `anova()` function in R to carry out this test.

Group-specific slopes: Interactions

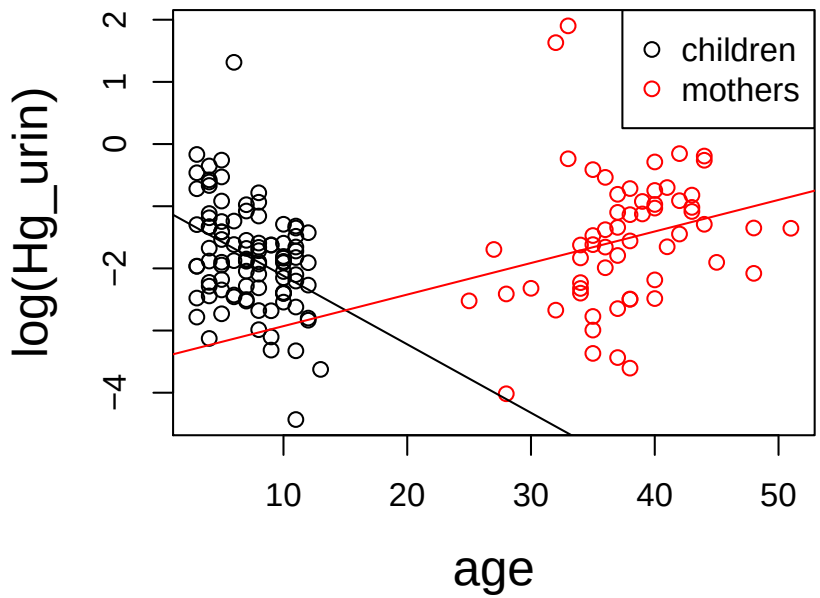
It may happen that groups do not only differ in their intercept (β_0), but also in their slopes (β_x).

In this case, one has to introduce an interaction term into the regression equation.

For simplicity, let us look at a binary covariate ($x_i \in \{0, 1\}$).
(Generalization to factorial covariates is more or less straightforward.)

Remember the mercury (Hg) example from last week. We now extended the dataset and include mothers *and* children ($\leq$ 11 years).

It is known that Hg concentrations may change over the lifetime of humans. So let us look at $\texttt{log}(\texttt{Hg}_{\texttt{urin}})$ depending on the participants age:



An important observation is that children and mothers show different dependencies of age!

It is therefore crucial to formulate a model that allows for different intercepts *and* slopes, depending on group membership (mother/child).

The smallest possible model is then given as

$$y_i = \beta_0 + \beta_1 \text{mother}_i + \beta_2 \text{age}_i + \beta_3 \text{age}_i \cdot \text{mother}_i + e_i \ , \qquad (1)$$

where $y_i = \log(Hg_{\text{urin}})_i$, and `mother` is a binary "dummy" variable that indicates if the person is a mother (1) or a child (0).

This results in essentially **two** models with group specific intercept and slope:

Mothers ($x_i = 1$): $\hat{y}_i = \beta_0 + \beta_1 + (\beta_2 + \beta_3)\text{age}_i$

Children ($x_i = 0$): $\hat{y}_i = \beta_0 + \beta_2 \text{age}_i$

Question: why is there a hat on $\hat{y}_i$ now? Difference to y_i?

Fitting model (1) in R is done as follows, where `age:mother` denotes the interaction term ($\text{age}_i \cdot \text{mother}_i$):

```
> r.hg <- lm(log(Hg_urin)~  mother + age + age:mother,d.hg)
> summary(r.hg)$coef
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -1.0188317 0.25250071 -4.034966 8.624100e-05
mother      -2.4176907 0.91198012 -2.651034 8.874694e-03
age         -0.1101447 0.03225589 -3.414715 8.188542e-04
mother:age   0.1609032 0.03965739  4.057333 7.912112e-05
```

Interpretation:

Mothers: $\hat{y}_i = -1.02 + (-2.42) + (-0.11 + 0.16) \cdot \text{age}_i$

Children: $\hat{y}_i = -1.02 + (-0.11) \cdot \text{age}$

- The Hg level drops in young children.
- The Hg level increases in adults (mothers).

On the previous slide we have actually fitted 2 models at the same time.

What is the advantage of this? Why is this usually better than fitting two separate models, one for children and one for mothers?

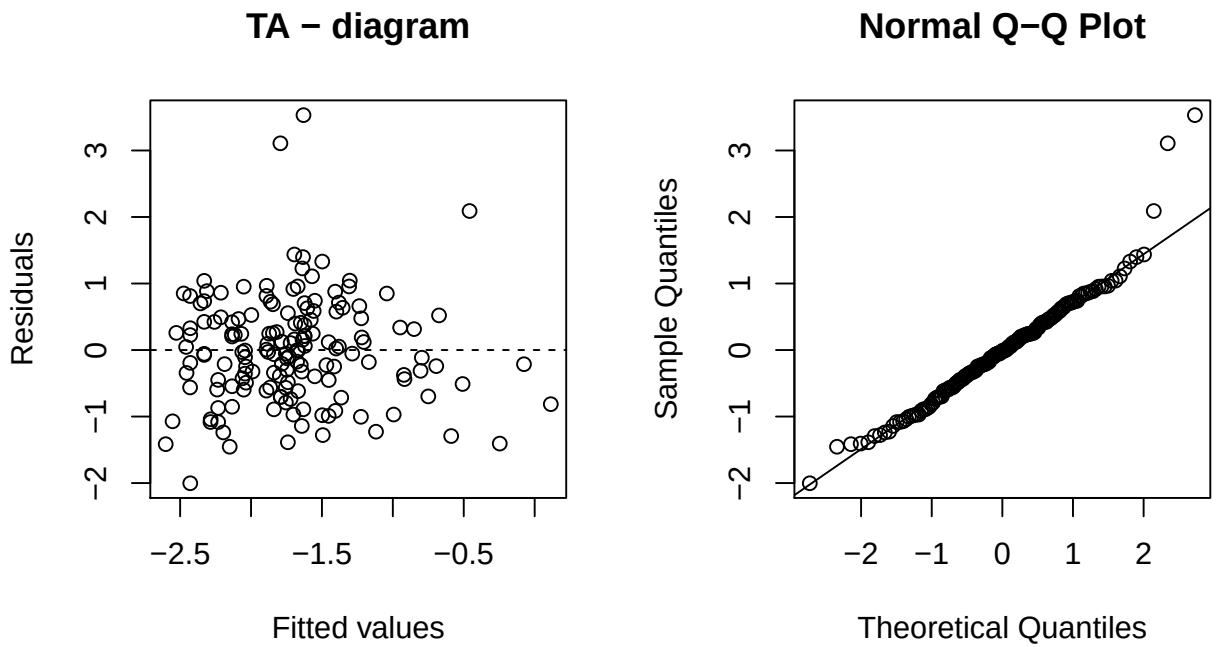
Remember (from last week), however, that the Hg model also included smoking status, amalgam fillings and fish consumption as important predictors. It is very straightforward to just include these predictors in model (1), which leads to the following model

```
> r.hg <- lm(log(Hg_urin)~  mother * age + smoking + amalgam + fish,d.hg)
```

	Coefficent	95%-confidence interval	p-value
Intercept	-1.35	from -1.82 to -0.87	< 0.0001
mother	-2.66	from -4.38 to -0.94	0.003
age	-0.098	from -0.16 to -0.04	0.001
smoking	0.60	from 0.06 to 1.15	0.03
amalgam	0.19	from 0.10 to 0.28	< 0.0001
fish	0.072	from 0.04 to 0.10	< 0.0001
mother:age	0.14	from 0.07 to 0.22	0.0001

(Note that `mother*age` in R encodes for `mother` + age + `mother:age`.)

Again, for completeness, some model checking (which one usually does before looking at the results):



Linear regression is even more powerful

We have seen that it is possible to include continuous, binary or factorial covariates in a regression model.

Even transformations of covariates be included in (almost) any form. For instance, include the square of a variable **x**

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i \ ,$$

which leads to a **quadratic** or **polynomial** regression (if higher order terms are used).

Other common transformations are (see also slide 37):

- log
- $\sqrt{..}$
- sin, cos,...

How can a *quadratic* regression be a *linear regression*??

Note: The word *linear* refers to the linearity in the coefficients, and not on a linear relationship between **y** and **x**!

Dieser Abschnitt hat gezeigt, dass das Modell der multiplen linearen Regression viele Situationen beschreiben kann, wenn man die X-Variablen geeignet wählt:

- Transformationen der X- (und Y-) Variablen können aus ursprünglich nicht-linearen Zusammenhängen lineare machen.
- Ein Vergleich von zwei Gruppen lässt sich mit einer zweiwertigen X-Variablen, von mehreren Gruppen mit einem „Block“ von dummy Variablen als multiple Regression schreiben. Auf diese Art werden nominale erklärende Variable in ein Regressionsmodell aufgenommen.
- Die Vorstellung von zwei verschiedenen Geraden für zwei Gruppen von Daten kann als ein einziges Modell hingeschrieben werden – das gilt auch für mehrere Gruppen. Auf allgemeinere Wechselwirkungen zwischen erklärenden Variablen kommen wir zurück (4.6.g).
- Die polynomiale Regression ist ein Spezialfall der multiplen linearen (!) Regression.

Multiple vs. many single regressions

Question: Given multiple regression covariates $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ...$ Could I simply fit separate simple models for each variable, that is

$y_i = \alpha + \beta x_i^{(1)} + e_i$

$y_i = \alpha + \beta x_i^{(2)} + e_i$

etc.?

Multiple vs. many single regressions

Question: Given multiple regression covariates $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, ...$ Could I simply fit separate simple models for each variable, that is

$y_i = \alpha + \beta x_i^{(1)} + e_i$

$y_i = \alpha + \beta x_i^{(2)} + e_i$

etc.?

Answer (Stahel 3.3o):

> Zusammenfassend: Ein multiples Regressionsmodell sagt mehr aus als viele einfache Regressionen – im Falle von korrelierten erklärenden Variablen sogar **viel mehr**.

Why?

Illustration

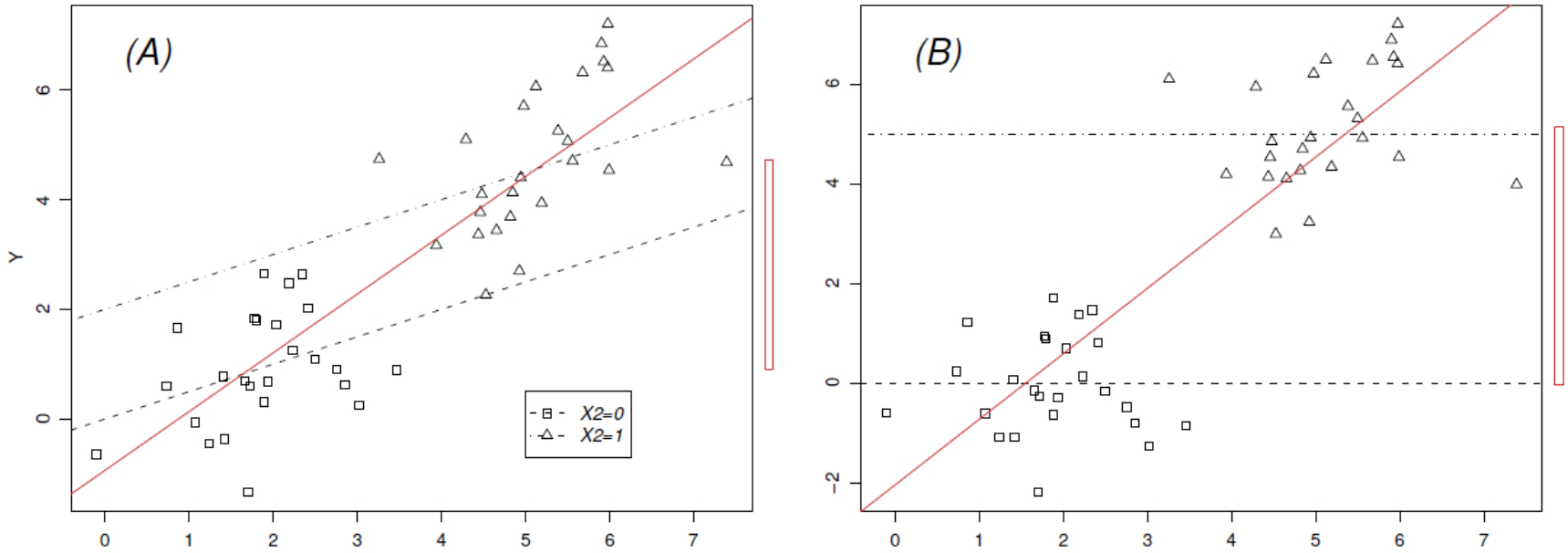
Chapter 3.3c in the Stahel script illustrates the point on four artificial examples. The model is always given as

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + e_i \ ,$$

where $\mathbf{x}^{(1)}$ is a continuous variable, and $\mathbf{x}^{(2)}$ is a binary grouping variable (thus taking values 0 for group 0 and 1 for group 1).

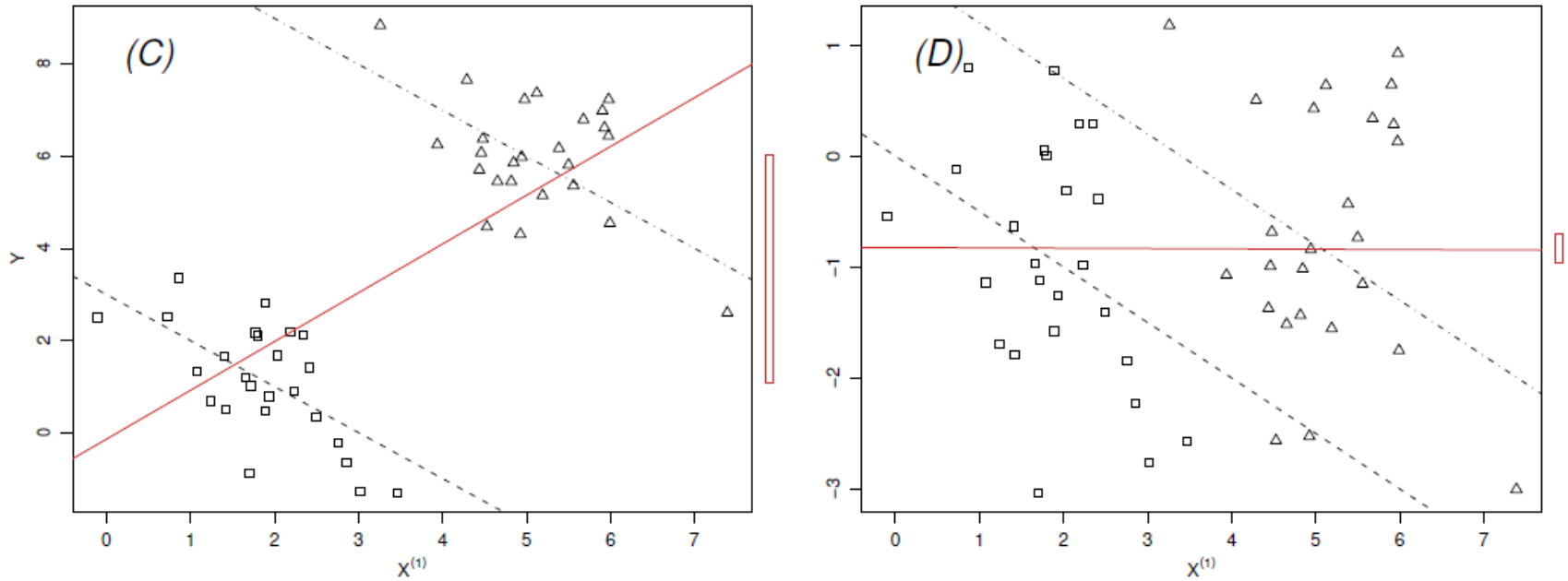
Thus the model is

$$\begin{aligned} \hat{y}_i &= \beta_0 + \beta_1 x_i^{(1)} && \text{if } x_i^{(2)} = 0. \\ \hat{y}_i &= \beta_0 + \beta_2 + \beta_1 x_i^{(1)} && \text{if } x_i^{(2)} = 1. \end{aligned}$$



Example A: Within-group slope is > 0. Fitting y against x leads to an overestimated slope when group-variable is not included in the model.

Example B: Within-group slope is 0, but fitting y against x leads to a slope estimate > 0, wich is only an artefact of not accounting for the group variable $x^{(2)}$.



Example C: Within-group slope is < 0, but fitting y against x leads to an estimated slope of > 0!

Example D: Within-group slope is < 0, but fitting y against x leads to a slope estimate of 0.

Another interpretation of multiple regression

In multiple regression, the coefficient β_x of a covariate x can be interpreted as follows:

> β_x explains how the response changes with x, while holding all the other variables constant.

This idea is similar in spirit to an experimental design, where the influence of a covariate of interest on the response is investigated in various environments[1]. Clayton and Hills (1993) continue (p.273):

> *To extend our analogy, the data analyst is in a position like that of an experimental scientist who has the capability to plan and carry out many experiments within a single day. Not surprisingly, a cool head is required!*

[1] Clayton, D. and M. Hills (1993). Statistical Models in Epidemiology. Oxford: Oxford University Press.

Checking modelling assumptions

Remember that in linear regression the modelling assumption is that the residuals e_i are independently normally distributed around zero, that is, $e_i \sim \mathrm{N}(0, \sigma_e^2)$. This implies four things:

a) The expected value of e_i is 0: $\mathrm{E}(e_i) = 0$.

b) All e_i have the same variance: $\mathrm{Var}(e_i) = \sigma_e^2$.

c) The e_i are normally distributed.

d) The e_i are independent of each other.

So far, we have discussed the Tukey-Anscombe plot and the QQ-plot.

Stahel 4.1b:

Diese Voraussetzungen zu überprüfen, ist meistens wesentlich. Es geht dabei nicht in erster Linie um eine Rechtfertigung, sondern um die Möglichkeit, aus allfälligen Abweichungen ein **besseres Modell** entwickeln zu können. Das kann bedeuten, dass

- Variable transformiert werden,
- zusätzliche Terme, beispielsweise Wechselwirkungen, ins Modell aufgenommen werden,
- für die Beobachtungen Gewichte eingeführt werden,
- allgemeinere Modelle und statistische Methoden verwendet werden.

The aim is to find a model that describes the data well. But always keep in mind the following statement from a wise man:

All models are wrong, but some are useful. (Box 1978)

Overview of model-checking tools

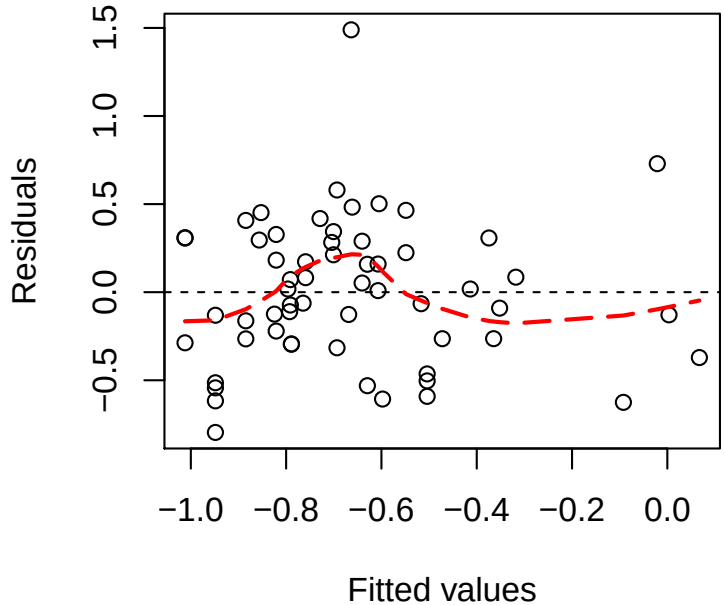
Complete overview of thos used in this course:

- Tukey-Anscombe plot (see lectures 2 and 3)
  $\Rightarrow$ To check assumptions a), b) and d)
- Quantile-quantile (QQ) plot (see lectures 2 and 3)
  $\Rightarrow$ To check assumption c)
- Scale-location plot (Streuungs-Diagramm)
  $\Rightarrow$ To check assumption b)
- Leverage plot (Hebelarm-Diagramm)
  $\Rightarrow$ To find influential observations and/or outliers

Note: these four diagrams are plotted automatically by R when you use the `plot()` or the `autoplot()` function (from the `ggfortify` package) on an `lm` object, for example `autoplot(r.hg)`.

Tukey-Anscombe plot

It is smometimes useful to enrich the TA-plot by adding a "running mean" or a "smoothed mean", which can give hints on the trend of the residuals. For the mercury example where $\log(Hg_{\text{urin}})$ is regressed on smoking, amalgam and fish consumption for mothers only (slides 32-34 of lecture 3):



The TA plot (again) indicates that there is a small problem in the range of -0.7 to -0.6, namely due to an outlier...

However, generally we recommend to not add a smoothing line, because it may bias our view on the plot.

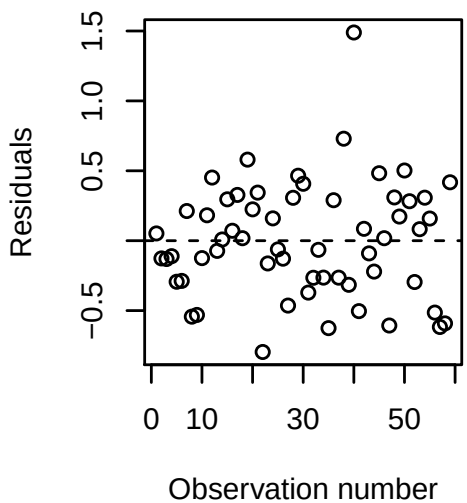
We claimed that the TA plot is also able to check the *independence assumption* d). But how?

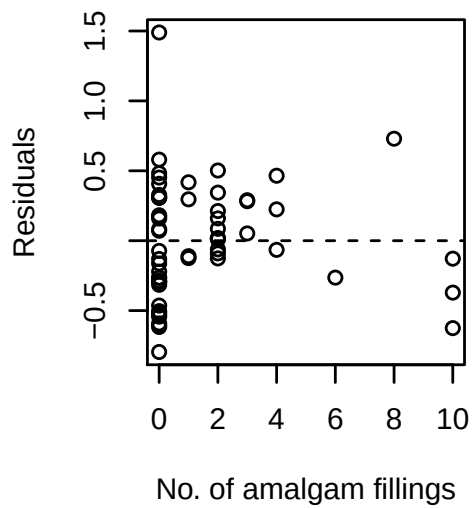
A dependency would be reflected by some kind of trend.

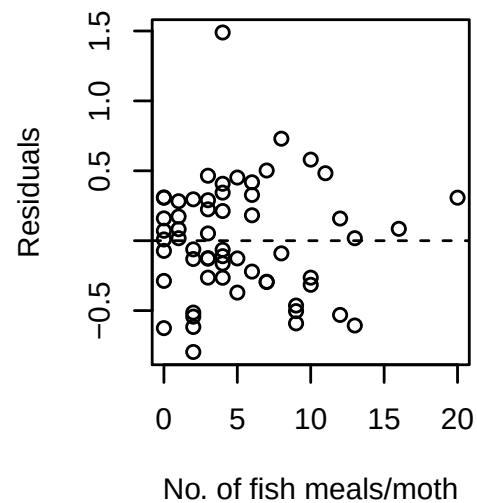
Other ideas to plot residuals to check for a dependency? Please discuss!

The dependency is not necessarily on the fitted values (x-axis of TA plot).
Ideas:

- Plot residuals in dependency of time (if available) or sequence of obervations.
- Plot residuals against the covariates.

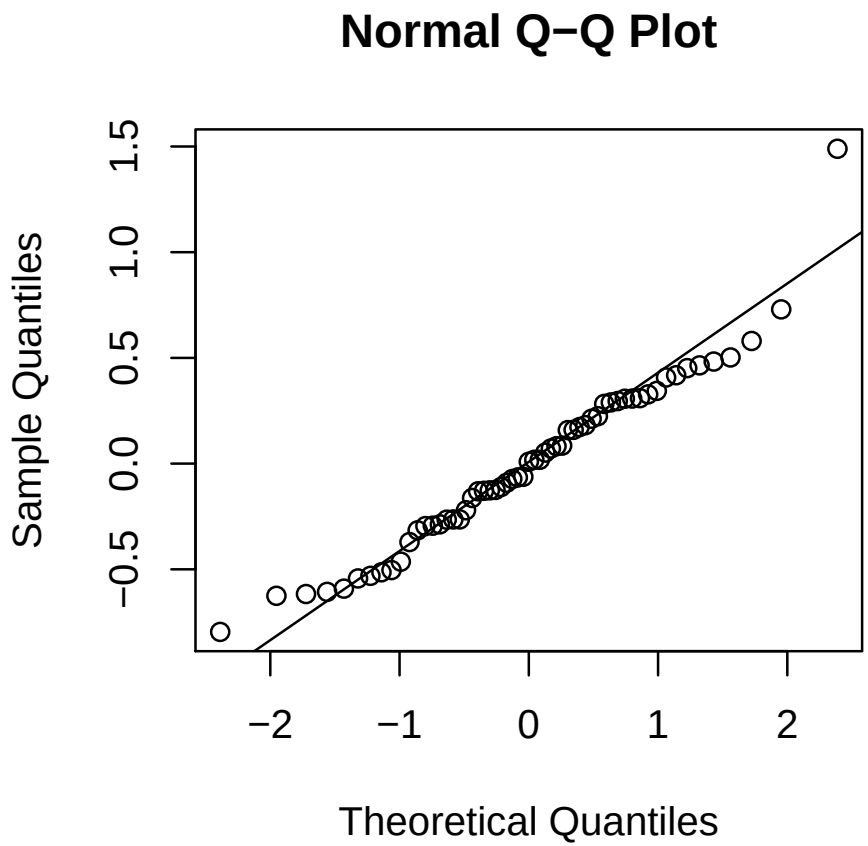






QQ-plot

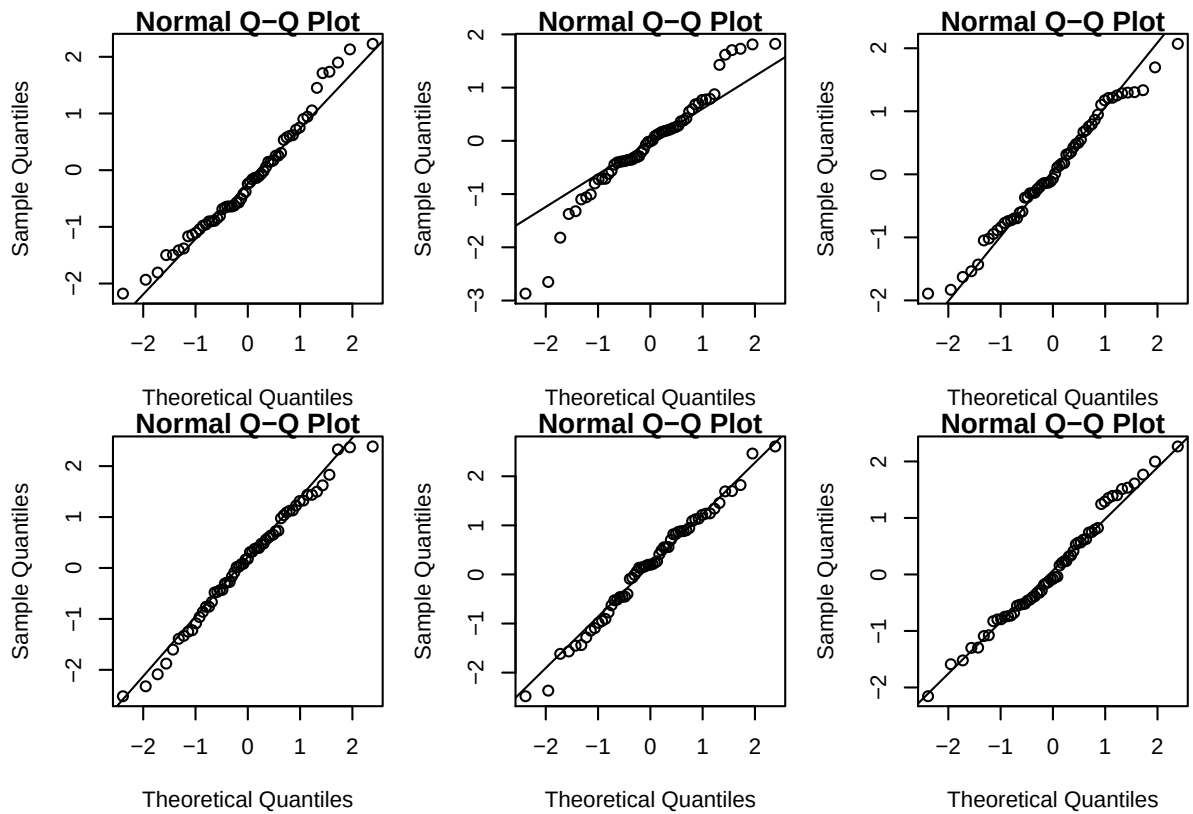
The outlier recorded above is also visible in the (well-known) QQ-plot, which is useful to check normal distribution of residuals (assumption c):



How do I know if a QQ-plot looks "good"?

There is no quantitative rule to answer this question, experience is needed. However, you can gain this experience from simulations. To this end, generate the same number of data points of a normally distributed variable and compare to your plot.

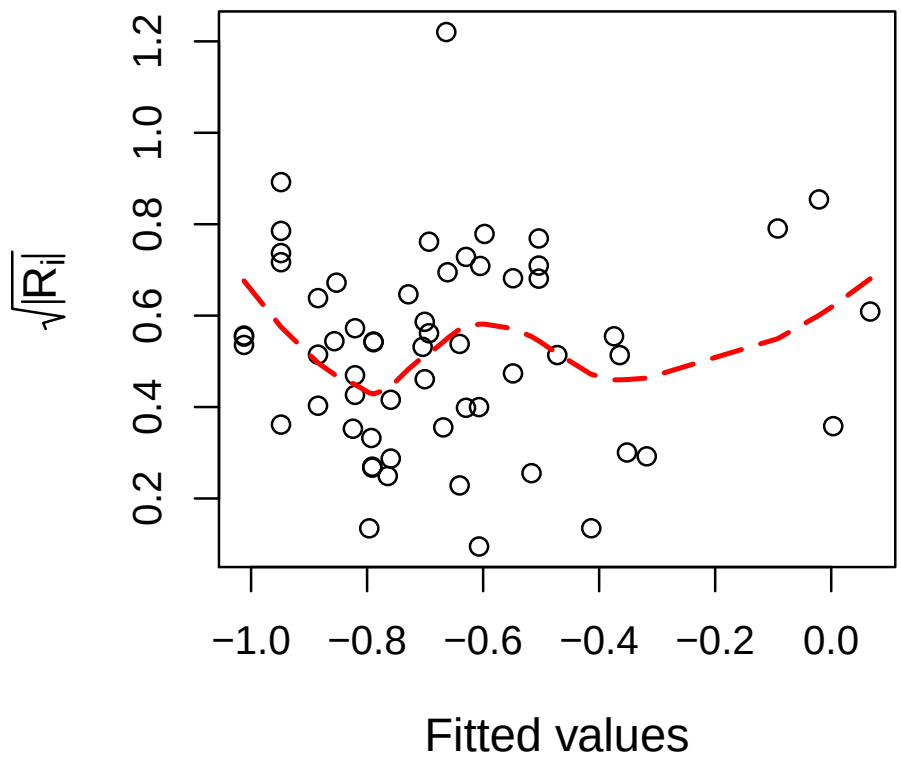
Example: Generate 59 points $e_i \sim N(0, 1)$ each time:



Scale-location plot (Streuungs-Diagramm)

The scale-location plot is particularly suited to check the assumption of equal variances (**homoscedasticity / Homoskedastizität**).

The idea is to plot the square root of the residuals $\sqrt{|R_i|}$ against the fitted values $\hat{y}_i$ (again using the Hg example with the mothers):



Leverages

To understand the leverage plot, we need to introduce the idea of the *leverage* ("Hebel"), see Stahel 4.3 h).

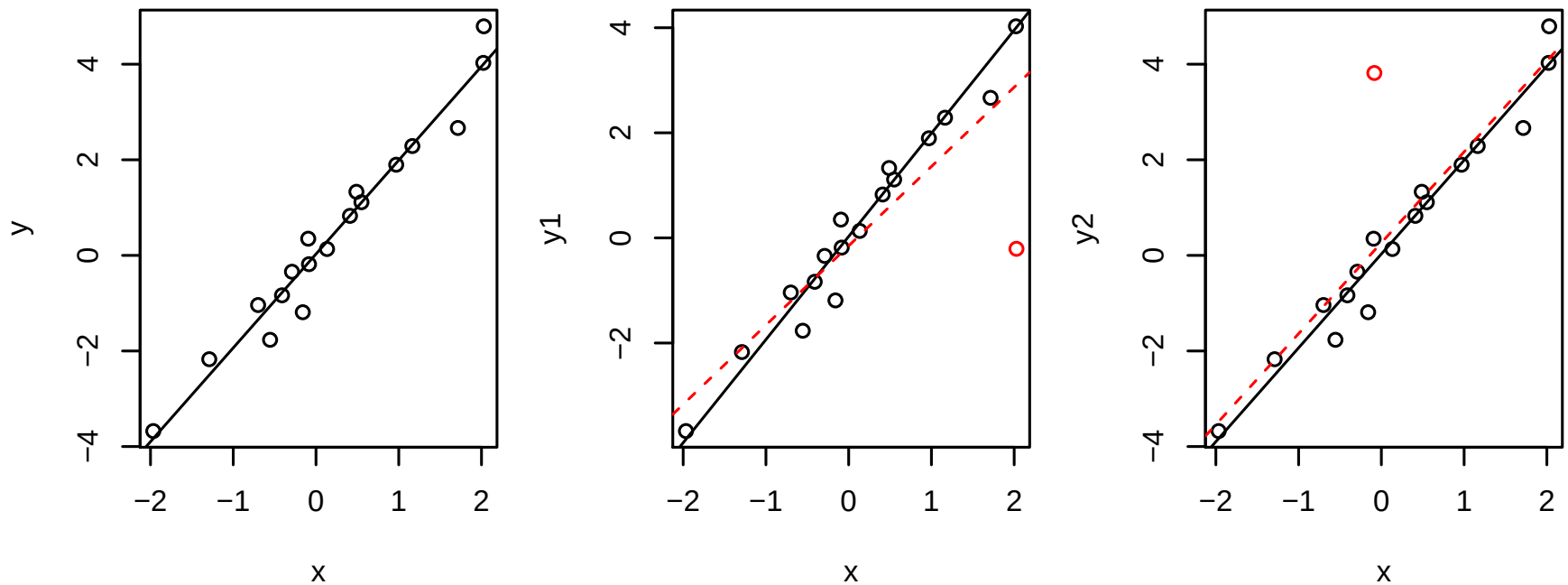
In simple regression, the leverage of individual i is defined as $H_{ii} = (1/n) + (x_i - \overline{x})^2 SSQ^{(X)}$, which becomes larger the further away from the mean...

h Die Hebelarm-Werte haben einige anschauliche Bedeutungen:

- Wenn man einen Wert Y_i um Δy_i verändert, dann misst $H_{ii}\Delta y_i$ die Veränderung des zugehörigen angepassten Wertes $\widehat{y}_i$. Wenn H_{ii} also gross ist, dann „zwingt die $i$te Beobachtung die Regressions-Funktion, sich an sie stark anzupassen". Sie hat eine „grosse **Hebelwirkung**" – daher der Name.
- Das macht auch das Ergebnis über die Varianzen qualitativ plausibel: Wenn die $i$te Beobachtung die Regressionfunktion stark an sich zieht, wird die Abweichung R_i tendenziell geringer, also die Varianz von R_i kleiner.
- Hebelpunkte in der Physik sind solche, die weit vom Drehpunkt entfernt sind. In unserem Zusammenhang heisst das, dass sie in gewissem Sinne weit vom „grossen Haufen" der Punkte weg sind, was die x-Variablen betrifft.

Graphical illustration of the leverage effect

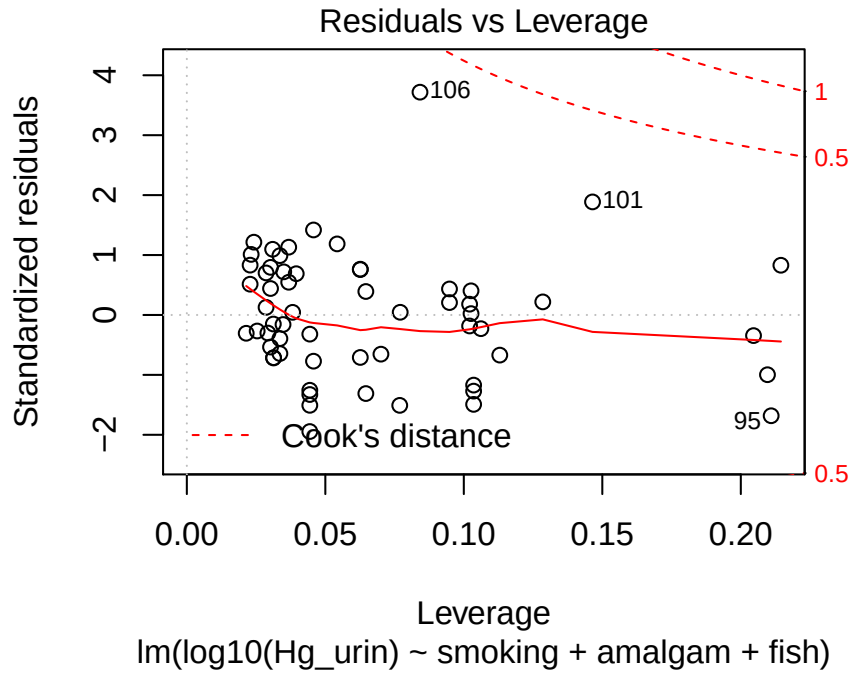
Data points with x_i values far from the mean have a stronger leverage effect than when $x_i \approx \overline{x}$:



The outlier in the middle plot "pulls" the regression line in its direction and biases the slope.

Leverage plot (Hebelarm-Diagramm)

In the leverage plot, (standardized) residuals $\tilde{R}_i$ are plotted against the leverage H_{ii} (continuing with the Hg example):



Note: Cook's distance measures how much the regression changes when the ith observation is omitted.

Critical ranges are the top and bottom right corners!!

Here, individuals 95, 101 and 106 are potential outliers.

What can go "wrong" during the modelling process?

- ...

What to do when things go wrong?

(i) Transform the outcome or the covariables.

(ii) Take care of outliers.

(iii) Use weighted regression (not discussed here).

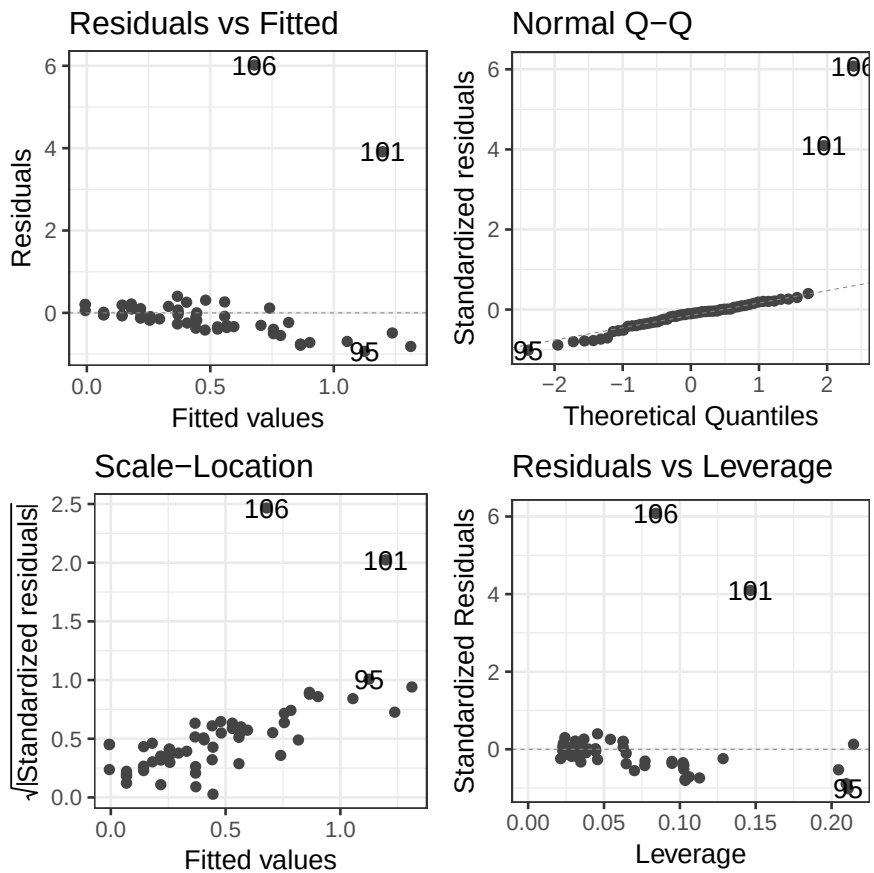
(iv) Improve the model, e.g., by adding additional terms or interactions (see "model selection" in week 7).

(v) Use another model family (generalized or nonlinear regression model).

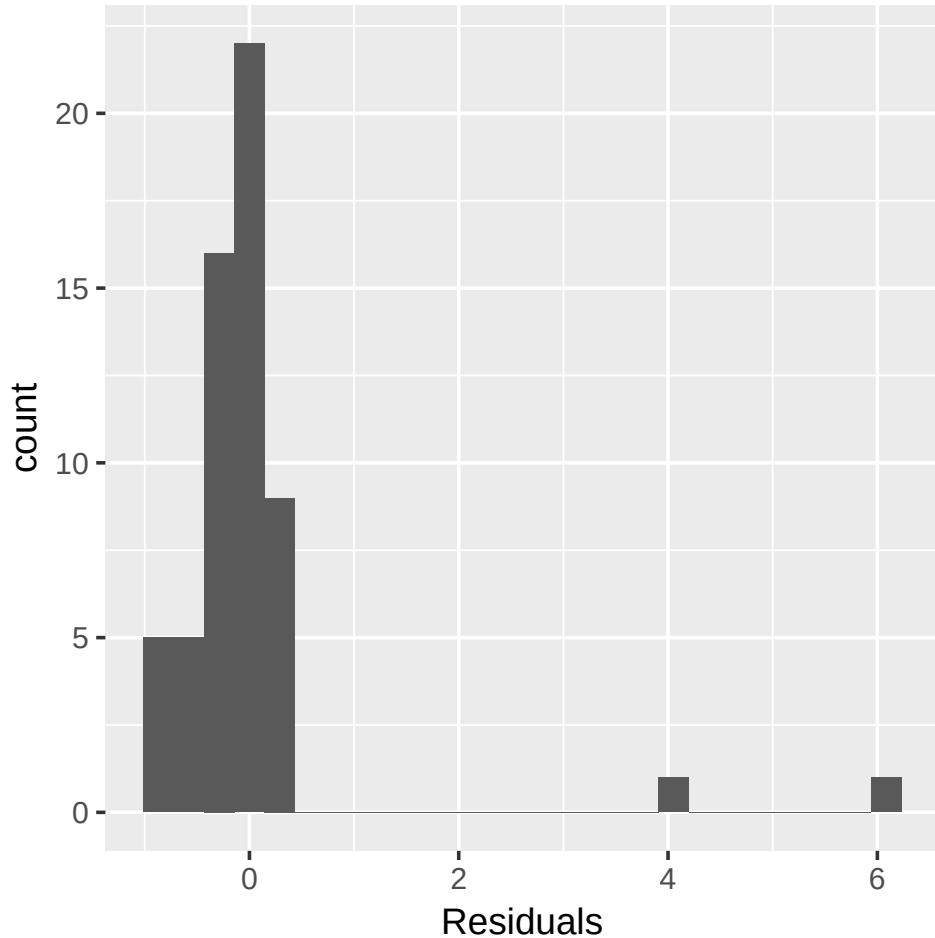
Transformation of the response?

Example: Use again the mercury study, include only mothers. Use the response (Hg-concentration in the urine) without log-transformation. What would it look like?

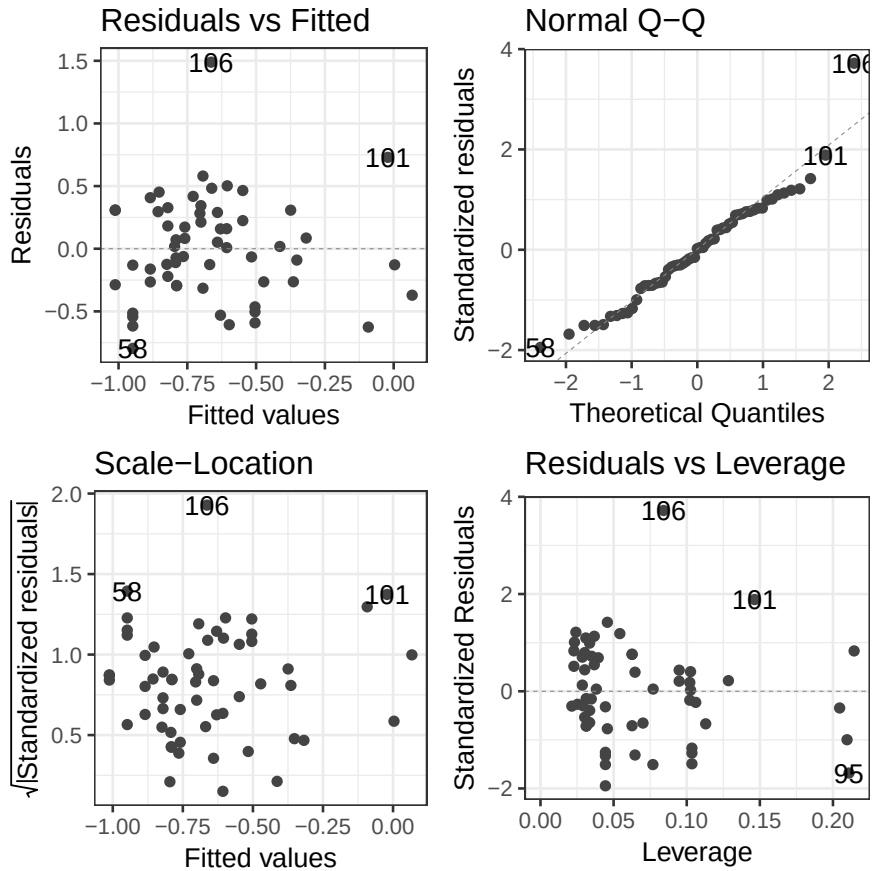
```
> r2.urin.mother <- lm(Hg_urin ~  smoking  + amalgam + fish,data=d.hg.m)
```



Also the "old-fashioned" histogram of the residuals is illustrative, it is very skewed:

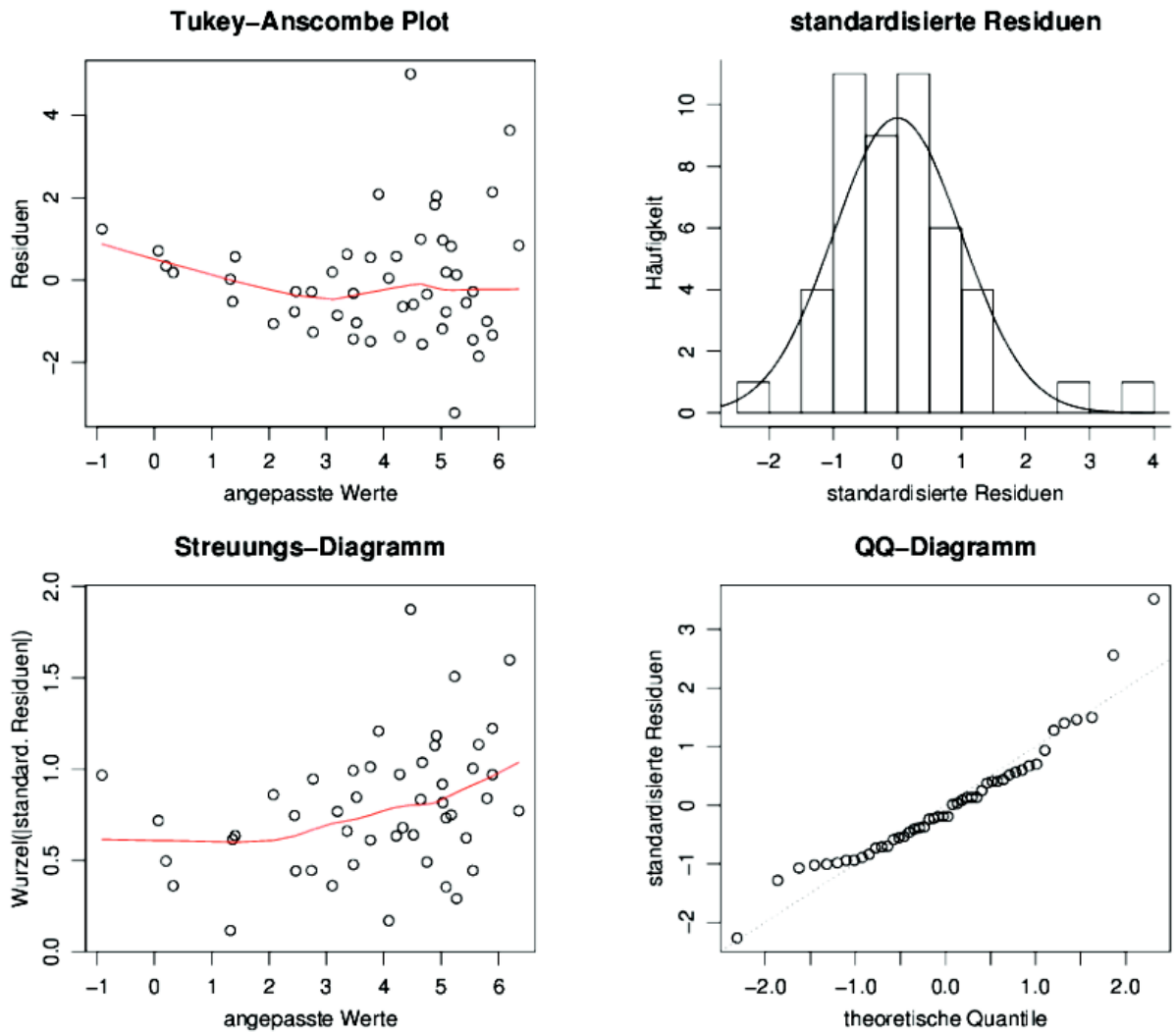


An in-summary comparison of the model with log-transformed response:



This looks **much** better! However... there is this individual 106 that needs some closer inspection (see slide 41 for the solution regarding this outlier).

A similar example is given in Stahel 4.4 a+b. The diagnostic plots of a regression where the log-transformation of the outcome was forgotten looked like this:



Common transformations

Which tranformations should be considered to cure model deviation symptoms? Answering this depends on plausibility and simplicity, and requires some experience.

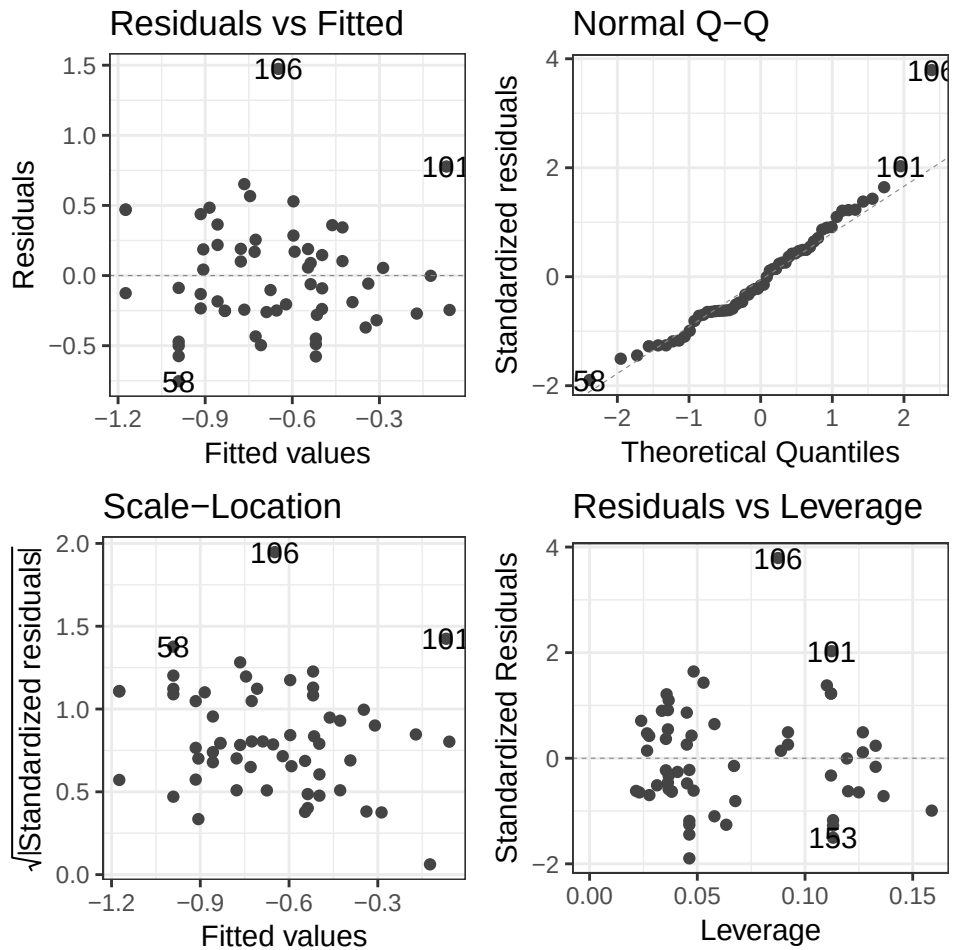
The most common and useful first aid transformations are:

- The log transformation for **concentrations** and **absolute values**.
- The square-root ($\sqrt{\cdot}$) transformation for **count data**.
- The arcsin($\sqrt{\cdot}$) transformation for **proportions/percentages**.

These transformations can (or should) also be applied on covariates!

For instance, the number of amalgam fillings and the number of monthly fish meals should be sqrt-transformed in the mercury example:

```
> r4.urin.mother <- lm(log10(Hg_urin) ~  smoking + sqrt(amalgam) + sqrt(fish),data=d.hg.m)
```



Outliers

(See Stahel chapter 4.5)

The above plots illustrate that outliers are visible in all diagnostic plots.

What to do in this case?

1. Start by checking the correctness of the data. Is there a typo or a digital point that was shifted by mistake? Check the covariates and the response.
2. If not, ask whether the model has been misspecified. Do reasonable transformations of the response or the covariables eliminate the outlier? Have the residuals a distribution with a long tail (which makes it more likely that extreme observations occur)?
3. Sometimes, an outlier may be the most interesting observation in a dataset!
4. Consider that outliers can also occur by chance!

Deleting outliers?

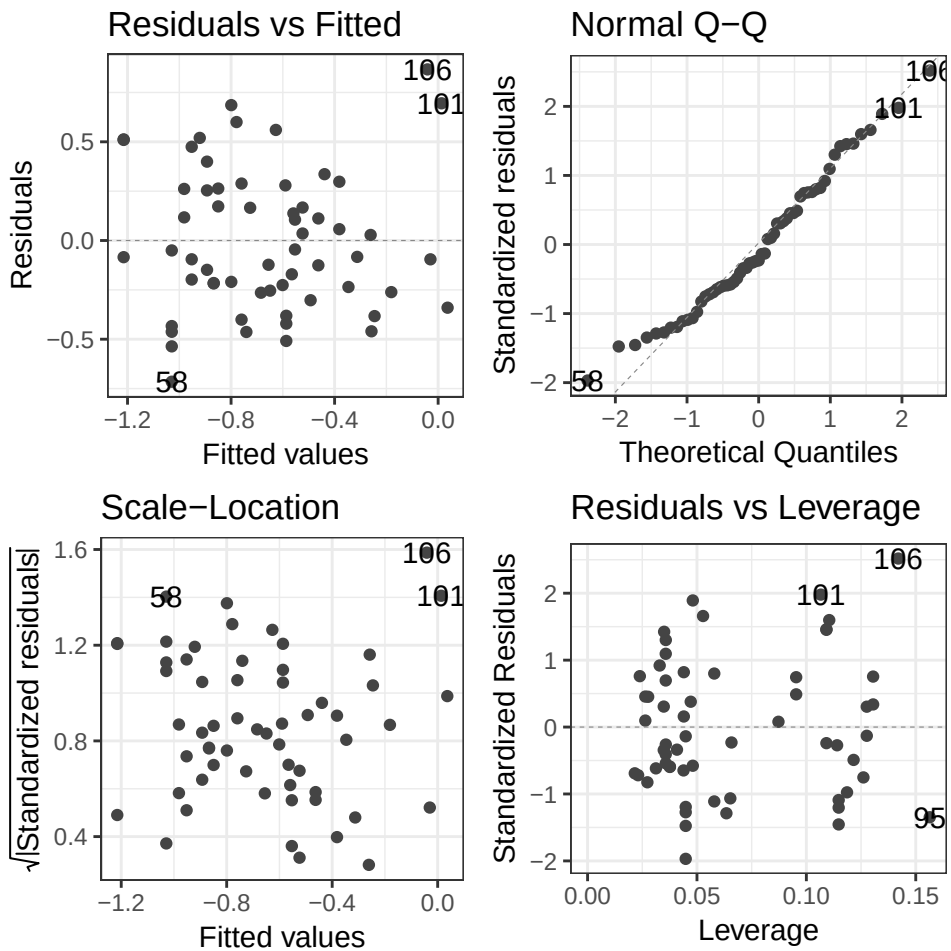
It might seem tempting to delete observations that apparently don't fit into the picture. However:

- Do this **only with absolute care**, e.g., if an observation has extremely implausible values!
- Before deleting outliers, check points 1-4 from the previous slide.
- When deleting outliers or the x% of most extreme observations, you **must mention this in your report**.
- Confidence intervals, tests and p-values might be biased.

The outlier in the Hg study

In the Hg study, it turned out later on that the outlier 106 had five unreported amalgam fillings!

A corrected analysis gives a much more regular picture (please compare to slide 38):



Summary