Lecture 5: Binary/categorical explanatory variables, and interactions

BIO144 Data Analysis in Biology

Stephanie Muff & Owen Petchey

University of Zurich

11 February, 2024

Overview

- Binary and categorical explanatory variables
- Interactions between explanatory variables
- Multiple vs. many single regressions
- Recap of checking model assumptions

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

- Intepretation of a regression model:
- How well does the model describe the data: Correlation and R^2
- Are the parameter estimates compatible with some specific value (t-test)?
- What range of parameters values are compatible with the data (confidence intervals)?
- What regression lines are compatible with the data (confidence band)?
- What are plausible values of other data (prediction band)?
- Multiple linear regression model $y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + \epsilon_i$.

Binary explanatory variables

So far, the explanatory variables x were always continuous.

In reality, there are **no restrictions assumed with respect to the x variables**.

One very frequent data type are **binary** variables, that is, variables that can only attain values 0 or 1.

See section 3.2c of the Stahel script:

If the binary variable x is the only variable in the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the model has only two predicted outcomes (plus error):

$$y_i = \begin{cases} \beta_0 + \epsilon_i & \text{if } x_i = 0 \\ \beta_0 + \beta_1 + \epsilon_i & \text{if } x_i = 1 \end{cases}$$

Sketch

Example: Smoking variable in Hg Study

For the 59 mothers in the Hg study, check if their smoking status (0=no,1=yes) influences the Hg-concentration in their urine.

We fit the following linear regression model:

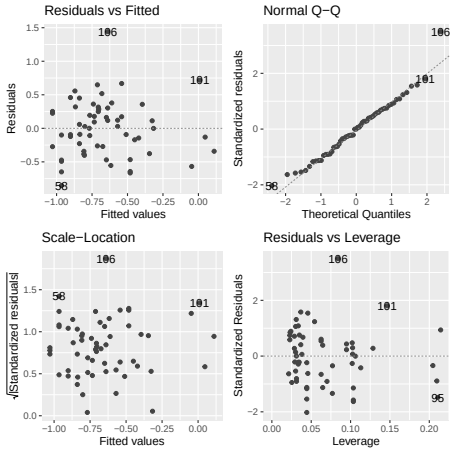
$$\log(Hg_{urin})_i = \beta_0 + \beta_1 \cdot x_i^{(1)} + \beta_2 \cdot x_i^{(2)} + \beta_3 \cdot x_i^{(3)} + \epsilon_i \ ,$$

Where

- $\log(Hg_{urin})$ is the urine mercury concentration.
- $x^{(1)}$ is the binary smoking indicator (0/1).
- $x^{(2)}$ the number of amalgam fillings.
- $x^{(3)}$ the monthly number of marine fish meals.

(Assume that we already looked at the data and see that log of Hg concentrations is needed.)

First check the modelling assumptions:



Seems ok, apart from one point (106) that could be categorized as an outlier. We ignore it for the moment.

The results table is given as follows:

	Coefficient	95%-confidence interval	p-value
Intercept	-1.03	from -1.24 to -0.81	< 0.0001
smoking	0.26	from -0.03 to 0.55	0.073
amalgam	0.098	from 0.05 to 0.14	< 0.0001
fish	0.032	from 0.01 to 0.06	0.019

There is some weak ($p = 0.073$) indication that smokers have an increased Hg concentration in their body. Their $\log(Hg_{urin})$ is in average by 0.26 (log10 units) higher than for nonsmokers.

In principle, we have now – at the same time – fitted **two models:** one for smokers and one for non-smokers, assuming that the slopes of the remaining explanatory variables are the same for both groups.

Smokers: $y_i = -1.03 + 0.26 + 0.098 \cdot amalgam_i + 0.032 \cdot fish_i + \epsilon_i$
Non-smokers: $y_i = -1.03 + 0.098 \cdot amalgam_i + 0.032 \cdot fish_i + \epsilon_i$

Categorical explanatory variables

Some explanatory variables indicate a **category**, for instance the species of an animal or a plant. This type of explanatory variable is termed **categorical**. For this there is trick: we can convert a categorical variable with k levels (for instance 3 species) into k dummy variables $x_i^{(j)}$ with

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

Each of the explanatory variables $x^{(1)}, \ldots, x^{(k)}$ can then be included as a binary variable in the model

$$y_i = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} + \epsilon_i \ .$$

However: this model is not identifiable. We could add a constant to $\beta_1, \beta_2, \ldots \beta_k$ and subtract it from β_0, and the model would fit equally well to the data, so it cannot be decided which set of the parameters is best.

Sketch (1)

Categorical explanatory variables (duplicate slide)

Some explanatory variables indicate a **category**, for instance the species of an animal or a plant. This type of explanatory variable is termed **categorical**. For this there is trick: we can convert a categorical variable with k levels (for instance 3 species) into k dummy variables $x_i^{(j)}$ with

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

Each of the explanatory variables $x^{(1)}, \ldots, x^{(k)}$ can then be included as a binary variable in the model

$$y_i = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} + \epsilon_i \ .$$

However: this model is not identifiable. We could add a constant to $\beta_1, \beta_2, \ldots \beta_k$ and subtract it from β_0, and the model would fit equally well to the data, so it cannot be decided which set of the parameters is best.

Solution...

Solution: One of the k categories must be selected as a *reference category* and is *included in the model as the intercept*. Typically: the alphabetically first category is the reference, thus $\beta_1 = 0$.

The model thus discriminates between the categories, such that (assuming $\beta_1 = 0$)

$$\hat{y}_i = \begin{cases} \beta_0, & \text{if } x_i^{(1)} = 1 \\ \beta_0 + \beta_2, & \text{if } x_i^{(2)} = 1 \\ \ldots & \\ \beta_0 + \beta_k, & \text{if } x_i^{(k)} = 1 \end{cases} .$$

Sketch (2)

Sketch (3)

!!Important to remember!!

(Common aspect that leads to confusion!)

Please note that a categorical variable with k categories requires $k-1$ parameters!

$\rightarrow$ The degrees of freedom are also reduced by $k-1$.

Degrees of freedom DF, example

- When we calculate something from the data and use it then we constrain the data $\rightarrow$ it has less "freedom"
- For example, to calculate $SS_{Tot} = \sum_{i=1}^{n}(y_i - \overline{y})^2$ we need $\overline{y}$
  - Now let's say we know that $\overline{y} = 4$ & $n = 5$, what values can y_i have?
  - 4 out of the 5 y_i can have whichever value, e.g. $y_i = \{2, 5, 3, 6, \square\}$, but given the first 4 values and given $\overline{y}$ & n, we know that the fifth value must be $\square = 5 \Rightarrow$ the fifth value is no longer free to vary
  - Thus, for the calculation of SS_{Tot} we have $n - 1$ degrees of freedom (DF), because we used 1 DF to calculate $\overline{y}$ from y_i (in other words, $\overline{y}$ and y_i are not independent)
- The DF are the number of data points that are free to vary, given what we want calculate
  - We lose one DF for every parameter that we want to estimate

Sketch (about degrees of freedom, 2)

Example: Earthworm study

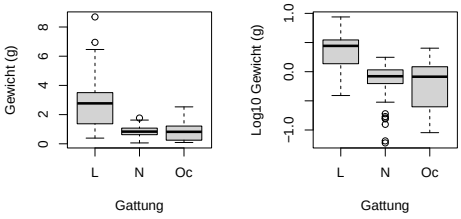
(Angelika von Förster und Burgi Liebst)

Die Dachse im Sihlwald ernähren sich zu einem grossen Prozentsatz von Regenwürmern. Ein Teil des Muskelmagens der Regenwürmer wird während der Passage durch den Dachsdarm nicht verdaut und mit dem Kot ausgeschieden. Wenn man aus der Grösse des Muskelmagenteilchens auf das Gewicht des Regenwurms schliessen kann, ist die Energiemenge berechnenbar, die der Dachs aufgenommen hat.

Frage: Besteht eine Beziehung zwischen dem Umfang des Muskelmagenteilchens und dem Gewicht des Regenwurms?

Data set of $n = 143$ worms with three species (Lumbricus, Nicodrilus, Octolasion), weight, stomach circumference (Magenumfang).

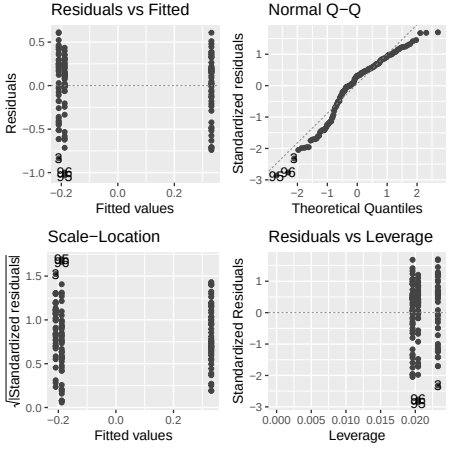
Data inspection suggests that the three species have different weights:



Formulate a model with $\log_{10}$(Gewicht) as response and `Gattung` as the explanatory variable.

```
r.lm <- lm(log10(GEWICHT) ~  Gattung, d.wurm)
```

Before doing anything else, check the modeling assumptions:



Results:

	Coefficient	95%-confidence interval	p-value
Intercept	0.33	from 0.23 to 0.43	< 0.0001
GattungN	-0.52	from -0.66 to -0.38	< 0.0001
GattungOc	-0.54	from -0.69 to -0.39	< 0.0001

$R^2 = 0.33$, $R^2_a = 0.32$.

- Question: Why is Gattung Lumbricus (L) not in the results table?
- Answer: L was chosen as the "reference category", thus $\beta_L = 0$.

Degrees of freedom: We had 143 data points. How many degrees of freedom are left for the residual error?

Interpreting the results I

- $\beta_0 = 0.33$ is the intercept.
- $\beta_2 = -0.52$ is the coefficient for Gattung=Nicodrilus.
- $\beta_3 = -0.54$ is the coefficient for Gattung =Octolasion.
- No coefficient is needed for Gattung Lumbricus, because $\beta_L = 0$.

We have now actually fitted **three** models, one model for each species:

Lumbricus: $\hat{y}_i = 0.33$

Nicodrilus: $\hat{y}_i = 0.33 + (-0.52)$

Octolasion: $\hat{y}_i = 0.33 + (-0.54)$

Interpreting the results III

Question: Is the "Gattung" explanatory variable relevant in the model, that is, do the model intercepts differ for the three species?

Problem: The *p*-values of the t-test for each of the worm species are not very meaningful. They belong to tests that compare the intercept of a category with the intercept of the reference level (i.e., the *difference* in intercepts!). However, the question is whether the variable `Gattung` has an effect in total.

Solution: To test if a categorical explanatory variable explains a significant amount of variability, we use an F-test (we saw this in regression; it is a "variance ratio" test.}

F-test

F-Test zum Vergleich von Modellen. Die Frage sei, ob die q Koeffzienten β_{j_1}, β_{j_2}, ..., β_{j_q} in einem linearen Regressionsmodell gleich null sein könnten.

- Nullhypothese: $\beta_{j_1} = 0$ und $\beta_{j_2} = 0$ und ... und $\beta_{j_q} = 0$
- Teststatistik:

$$T = \frac{(\mathrm{SSQ}^{(E)*} - \mathrm{SSQ}^{(E)})/q}{\mathrm{SSQ}^{(E)}/(n-p)};$$

  $\mathrm{SSQ}^{(E)*}$ ist die Quadratsumme des Fehlers im „kleinen“ Modell, die man aus einer Regression mit den verbleibenden $m-q$ X-Variablen erhält, und p die Anzahl Koeffizienten im „grossen“ Modell ($= m+1$, falls das Modell einen Achsenabschnitt enthält, $= m$ sonst).
- Verteilung von T unter der Nullhypolthese: $T \sim \mathcal{F}_{q,n-p}$, F-Verteilung mit q und $n-p$ Freiheitsgraden.

Der Test heisst F-Test zum Vergleich von Modellen. Allerdings kann nur ein kleineres Modell mit einem grösseren verglichen werden, in dem alle X-Variablen des kleinen wieder vorkommen, also mit einem „umfassenderen“ Modell. Der früher besprochene F-Test für das gesamte Modell (3.1.e) ist ein Spezialfall: das „kleine“ Modell besteht dort nur aus dem Achsenabschnitt β_0.

F-test (sketch 1)

F-test (sketch 1)

F-test for the earthworms

The function `anova()` in R does the F-test for categorical variables.

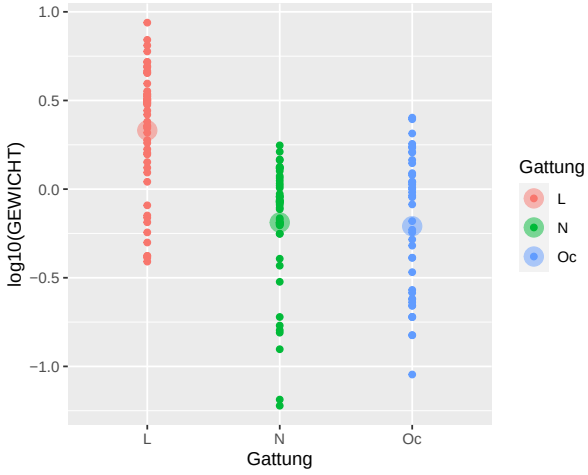
```
anova(r.lm)
```

```
## Analysis of Variance Table
##
## Response: log10(GEWICHT)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Gattung     2  9.2044  4.6022  34.568 6.293e-13 ***
## Residuals 140 18.6388  0.1331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: Here, the F-value for `Gattung` is distributed as $F_{2,140}$ under the Null-Hypothesis.

This gives $p = 6.293e - 13$, thus a clear difference in the regression models for the three species ("Gattung is relevant").

Plotting the earthworms results

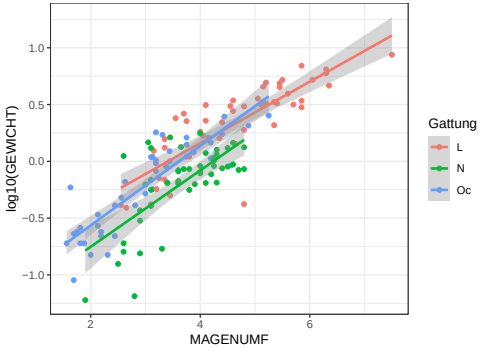


That was a lot...

Binary and categorical variables...

- Take a few minutes to consolidate, identify questions, ask them.

The earthworm data has more to learn from



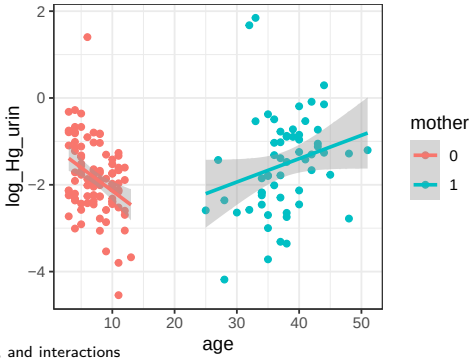
This model will be fitted in this week's BC videos.

Binary variable with interaction

For simplicity, let us look at a binary explanatory variable ($x_i \in \{0, 1\}$).

Remember the mercury (Hg) example. We now extended the dataset and include mothers **and** children (≤ 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 age of the participants:



Observation: **The regression lines are not parallel.**

→ Children and mother's Hg level seem to depend differently on age!

What does this mean for the model?

→ Formulate a model that allows for different intercepts *and* slopes, depending on group membership (mother/child).

→ This can be achieved by introducing a so-called interaction term into the regression equation.

The smallest possible model is then given as

$$y_i = \beta_0 + \beta_1 \text{mother}_i + \beta_2 \text{age}_i + \beta_3 \text{age}_i \cdot \text{mother}_i + \epsilon_i \ , \tag{1}$$

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

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, data = d.hg)
summary(r.hg)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -1.0685772 0.25997341 -4.110333 6.442221e-05
## mother1     -2.4811450 0.93896995 -2.642411 9.093456e-03
## age         -0.1071729 0.03321050 -3.227078 1.531867e-03
## mother1:age  0.1609570 0.04083104  3.942026 1.229208e-04
```

Interpretation:

Mothers: $\hat{y}_i = -1.06 + (-2.48) + (-0.10 + 0.16) \cdot \text{age}_i$

Children: $\hat{y}_i = -1.06 + (-0.10) \cdot \text{age}$

- 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?

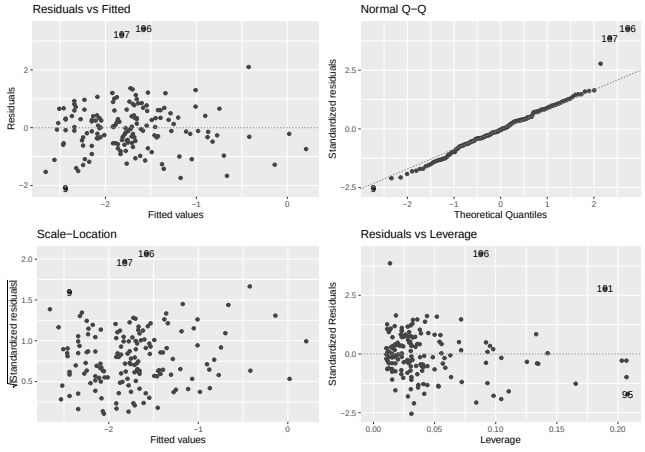
Remember, however, that the Hg model also included smoking status, amalgam fillings and fish consumption as important predictors. It is very straightforward to just include these predictors in model (1), which leads to the following model

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

	Coefficient	95%-confidence interval	p-value
Intercept	-1.38	from -1.87 to -0.89	< 0.0001
mother1	-2.75	from -4.52 to -0.98	0.003
age	-0.096	from -0.16 to -0.04	0.002
smoking	0.70	from 0.14 to 1.26	0.015
amalgam	0.20	from 0.11 to 0.29	< 0.0001
fish	0.069	from 0.04 to 0.10	< 0.0001
mother1:age	0.14	from 0.07 to 0.22	0.0002

(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 categorical explanatory variable in a regression model.

Even transformations of explanatory variables can be included in (almost) any form. For instance the square of a variable x

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

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

Other common transformations are (see also slide 38):

- 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 variables $x^{(1)}, x^{(2)}, \ldots$. Could I simply fit separate simple models for each variable, that is

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \epsilon_i$$

$$y_i = \beta_0 + \beta_1 x_i^{(2)} + \epsilon_i$$

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**.

Another interpretation of multiple regression

In multiple regression, the coefficient β_x of explanatory variable 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 an explanatory variable of interest on the response is investigated in various environments[1]. Clayton and Hills (1993) continue (p.273):

> *[...] 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 modeling assumptions (recap)

Remember that in linear regression the modeling assumption is that the errors ϵ_i are independently normally distributed around zero, that is, $\epsilon_i \sim N(0, \sigma^2)$. This implies four things:

a) The expected value of each residual ϵ_i is 0: $E(\epsilon_i) = 0$.
b) All ϵ_i have the same variance: $Var(\epsilon_i) = \sigma^2$.
c) The ϵ_i are normally distributed.
d) The ϵ_i are independent of each other.

The aim is to formulate 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

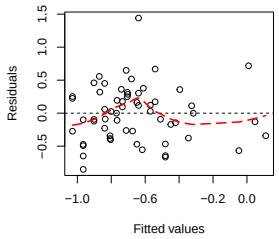
Overview of tools used in this course:

- Tukey-Anscombe plot (see lectures 3 and 4)
- Quantile-quantile (QQ) plot (see lectures 3 and 4)
- Scale-location plot (Streuungs-Diagramm)
- Leverage plot (Hebelarm-Diagramm)

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 sometimes 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:



The TA plot (again) indicates that there is an outlier in the range of -0.7 to -0.6.

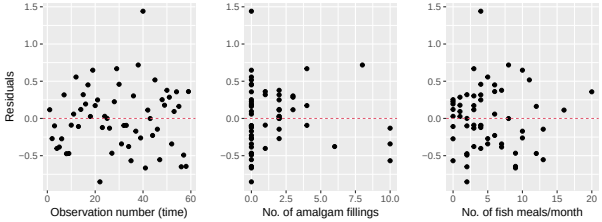
However, generally we recommend to not add a smoothing line, because it may bias our view on the plot.

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

$\rightarrow$ A dependency would be reflected by some kind of trend.

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

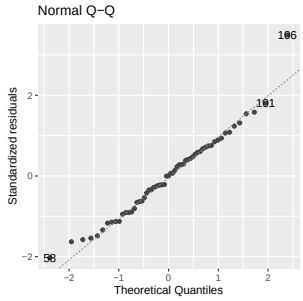
- Plot residuals in dependency of time (if available) or sequence of obervations.
- Plot residuals against the explanatory variable.



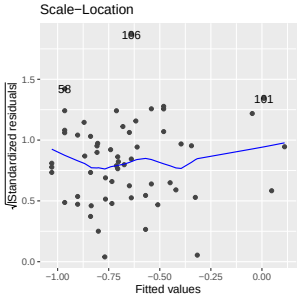
Again, no pattern = good.

QQ-plot

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



Check equal variance assumption



Leverage plot (Hebelarm-Diagramm)

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

Critical ranges are the top and bottom right corners!!

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

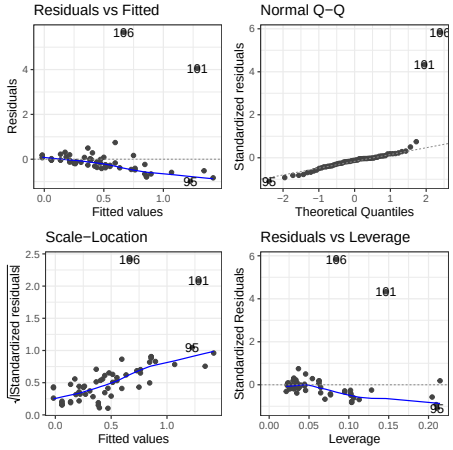
What to do when things go wrong?

1. Transform the response and/or explanatory variables.
2. Take care of outliers.
3. Use weighted regression (not discussed here).
4. Improve the model, e.g., by adding additional terms or interactions (see "model selection" in lecture 8).
5. Use another model family (generalized or nonlinear regression model).

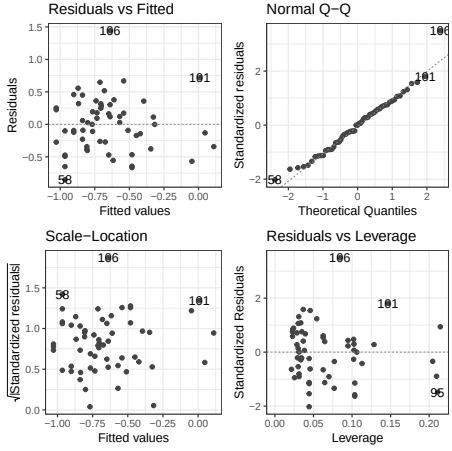
Transformation of the response?

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)
```



Comparison to the model with log-transformed response:



This looks **much** better! However... there is this individual 106 that needs some closer inspection (see slide 43 for the solution regarding this outlier).

Common transformations

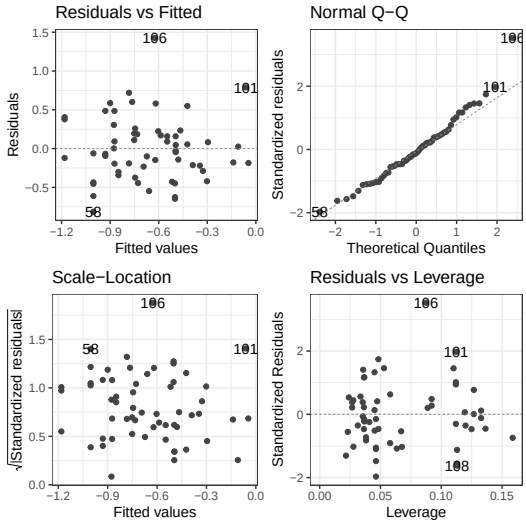
Which transformations should be considered to cure model deviation symptoms?
Answering this depends on plausibility and simplicity, and requires some experience.

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 also be applied on explanatory variables!

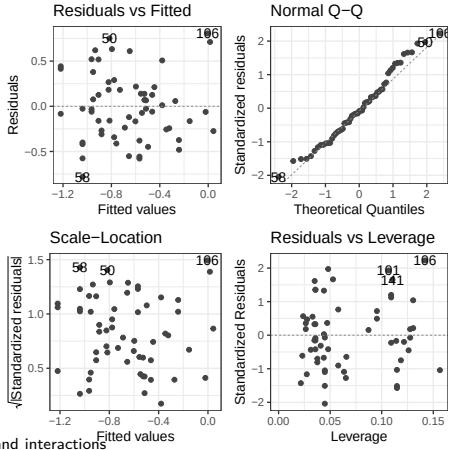
For instance, the number of amalgam fillings and the number of monthly fish meals could be sqrt-transformed in the mercury example:



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 40):



Recap

- Binary and categorical explanatory variables.
- Interactions: a categorical explanatory variables allows for group-specific intercepts and slopes (see earthworm example).
- The *F*-test is used to test if $\beta_2 = \beta_3 = ... = \beta_k = 0$ at the same time for a categorical explanatory variable with k levels. Use the `anova()` function in R to carry out this test.
- The *F*-test is a generalization of the *t*-test, because the latter is used to test $\beta_j = 0$ for one single variable $x^{(j)}$.
- Test for a single $\beta_j = 0 \rightarrow$ *t*-test.
- Test for several $\beta_2 = ... = \beta_k = 0$ simultaneously $\rightarrow$ *F*-test.

Thus you will **always** need the *F*-test `anova()` to obtain a *p*-value for a categorial explanatory variable with more than 2 levels!

Next steps

- (BC) Before (practical) Class: three videos (20, 13, and 18 mins) going through analysis of the earthworm study data.
- (IC) In (practical) Class: naked mole rats, reaction times (more independent)
- Then week 6: Analysis of variance (ANOVA).