

Lecture 5: Multiple linear regression (finalize) / Residual analysis / Checking modeling assumptions

BIO144 Data Analysis in Biology

Stephanie Muff & Owen Petchey

University of Zurich

17 March, 2021

Overview

- ▶ Interactions between explanatory variables
- ▶ 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*

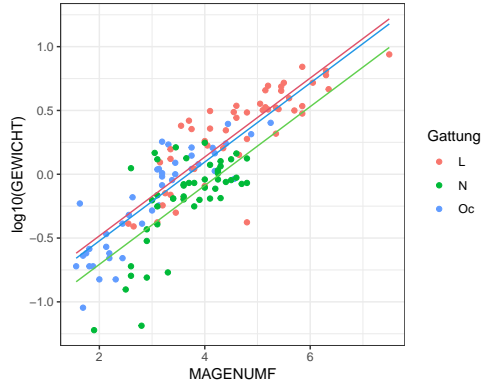
Recap of last week

- ▶ Multiple linear regression model $y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_m x_i^{(m)} + \epsilon_i$.
- ▶ **Binary** and **categorical** explanatory variables: The idea is to 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 $x^{(2)}, \dots, x^{(k)}$ in the regression, given that $x^{(1)}$ is used as **reference category** ($\beta_1 = 0$).
- ▶ The categorical explanatory variables of last week were used to allow for **group-specific intercepts** (see earthworm example).

Recap of last week II



- The **F-test** is used to test if $\beta_2 = \beta_3 = \dots = \beta_k = 0$ at the same time for a categorical explanatory variable with k levels. Use the `anova()` function in R to carry out this test.

Recap of last week III

- ▶ The F -test is a **generalization of the t -test**, because the latter is used to test $\beta_j = 0$ for one single variable $x^{(j)}$.

Cooking rule:

- ▶ Test for a single $\beta_j = 0 \rightarrow t$ -test.
- ▶ Test for several $\beta_2 = \dots = \beta_k = 0$ simultaneously $\rightarrow F$ -test.

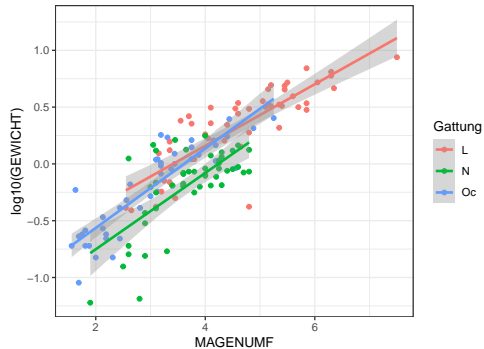
\rightarrow `anova()`

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

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 the earthworm example, allowing for different intercepts and slopes:



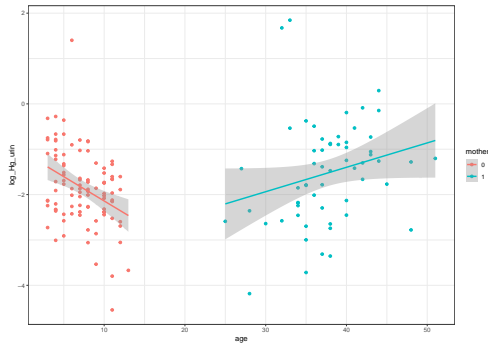
Important: This model will be fitted in this week's BC videos.

Binary variable with interaction

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

Remember the mercury (Hg) example from last week. 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 $\log(\text{Hg}_{\text{urin}})$ 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

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

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?

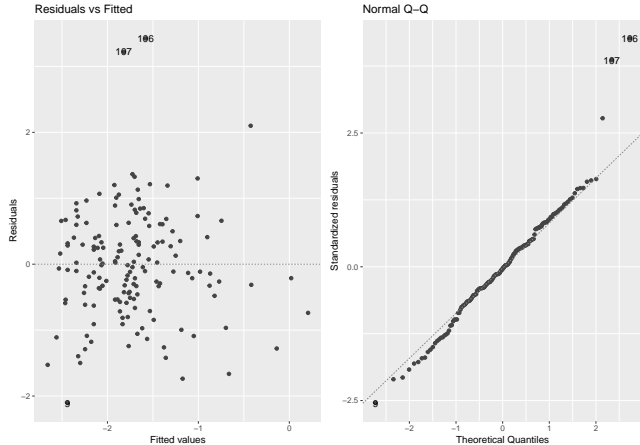
Remember (from last week), however, that the Hg model also included smoking status, amalgam fillings and fish consumption as important predictors. It is very straightforward to just include these predictors in model (1), which leads to the following model

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

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

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

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

etc.?

Multiple vs. many single regressions

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

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

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

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 environments¹. 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!

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

Checking modeling assumptions

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.

So far, we have discussed

- ▶ the Tukey-Anscombe plot.
- ▶ the QQ-plot.

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

Complete overview of tools used in this course:

- ▶ Tukey-Anscombe plot (see lectures 3 and 4)

⇒ To check assumptions a), b) and d)

- ▶ Quantile-quantile (QQ) plot (see lectures 3 and 4)

⇒ To check assumption c)

- ▶ Scale-location plot (Streuungs-Diagramm)

⇒ To check assumption b)

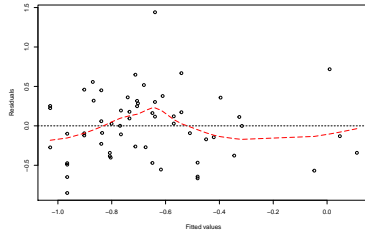
- ▶ Leverage plot (Hebelarm-Diagramm)

⇒ 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 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 for mothers only (slides 32-34 of lecture 4):



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

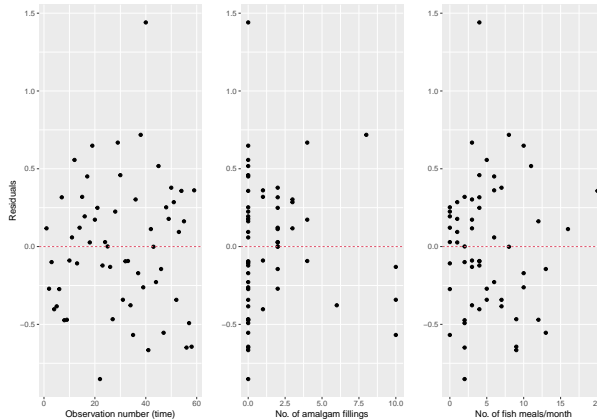
However, generally we recommend to **not** add a smoothing line, because it may bias our view on the plot.

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

→ A dependency would be reflected by some kind of **trend**.

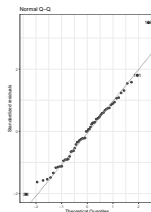
But: The dependency is not necessarily on the fitted values (x-axis of TA plot) Ideas:

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



Again, no pattern = good.

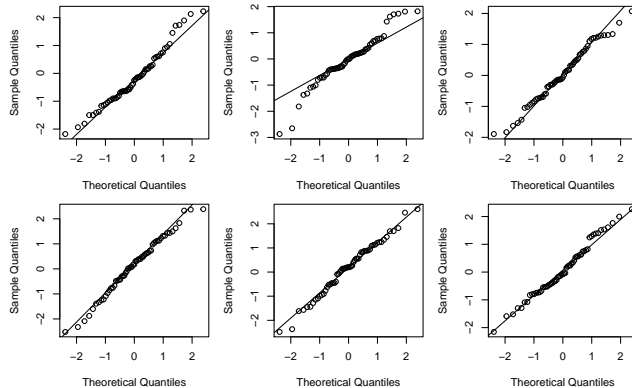
The **outlier** recorded above is also visible in the QQ-plot, which is useful to check for normal distribution of residuals (assumption c):



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.

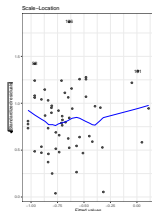
Example: Generate 59 points $\epsilon_j \sim N(0, 1)$ each time:



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 (standardized) residuals $\sqrt{|R_i|}$ against the fitted values \hat{y}_i . There should be **no trend** (Again Hg example):



Sketch explanation of scale-location

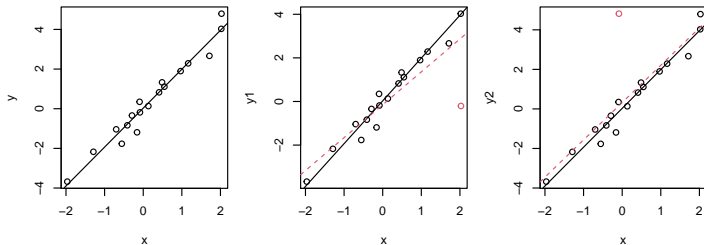
Leverages (“Hebel”)

To understand the leverage plot, we need to introduce the idea of the *leverage* (“Hebel”).

In simple regression, the leverage of individual i is defined as $H_{ii} = (1/n) + (x_i - \bar{x})^2 / SSQ^{(X)}$. Think about when leverages are expected to be large/small, and answer the two questions here:

Graphical illustration of the leverage effect

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

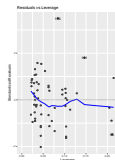


The outlier in the middle plot “pulls” the regression line in its direction and biases the slope.

→ [Click here to do it manually!](#)

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 can go “wrong” during the modeling process?



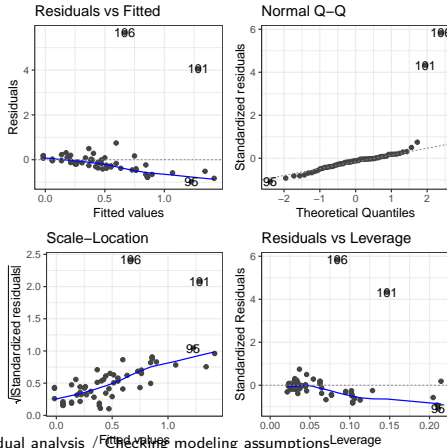
What to do when things go wrong?

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).

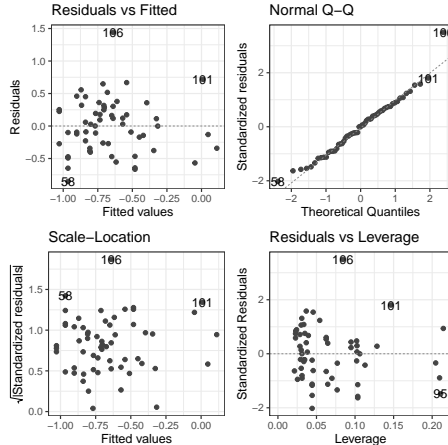
Transformation of the response?

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

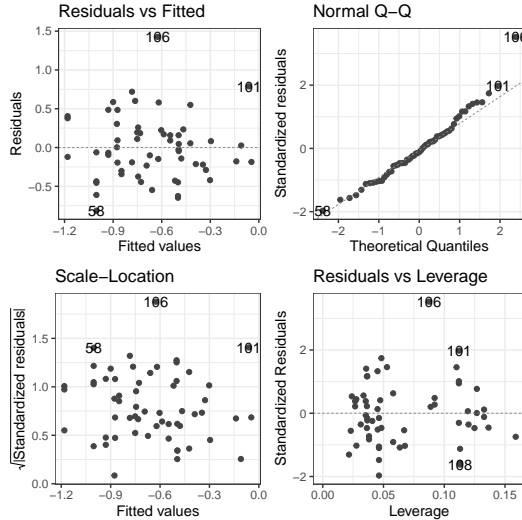
Which transformations should be considered to cure model deviation symptoms?
Answering this depends on plausibility and simplicity, and requires some experience.

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:



Outliers

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 both the response and explanatory variables.
2. If not, ask whether the model has been misspecified. Do reasonable transformations of the response and/or explanatory variables eliminate the outlier? Do the residuals have a distribution with a long tail (which makes it more likely that 3. 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!
5. Check in inclusion/exclusion changes the qualitative conclusion, and by how much it changes the quantitative conclusion.

Deleting outliers

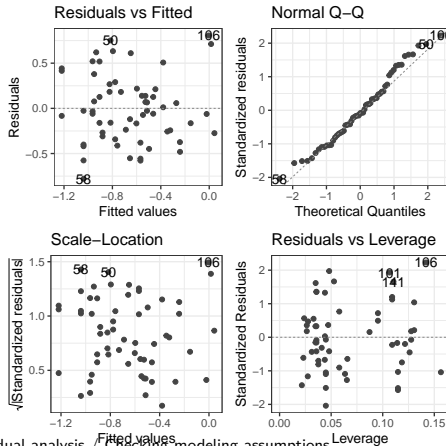
It might seem tempting to delete observations that apparently don't fit into the picture. However:

- ▶ 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**.

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):



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).