Kurs Bio144: Datenanalyse in der Biologie

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

Stefanie Muff (Lecture) & Owen L.Petchey (Practical)

University of Zurich

31 December, 2020

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*

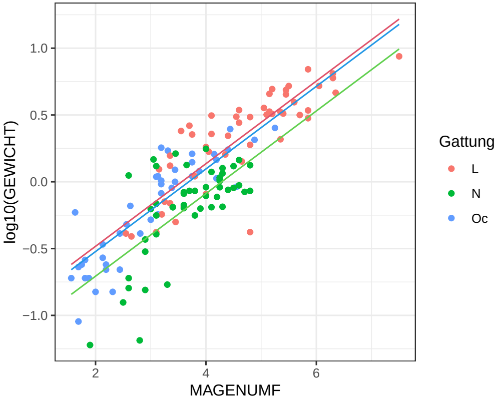
Recap of last week

- Multiple linear regression model $y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + \epsilon_i$.
- Binary and factor covariates: 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 $x^{(2)}, \ldots \, ,x^{(k)}$ in the regression, given that $x^{(1)}$ is used as reference category ($\beta_1 = 0$).
- The factor covariates 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 $\beta_2 = \beta_3 = ... = \beta_k = 0$ at the same time for a factor covariate with k levels. Use the `anova()` function in R to carry out this test.

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 $\beta_2 = ... = \beta_k = 0$ simultaneously $\rightarrow F$-test.

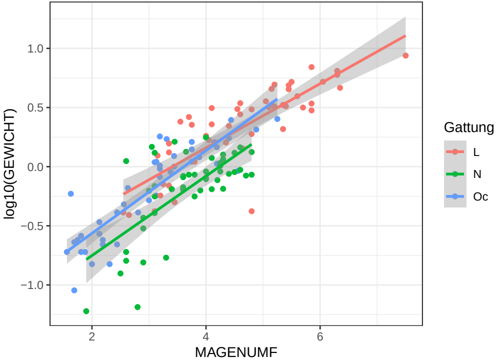
$\rightarrow$ `anova()`

Thus you will **always** need the F-test `anova()` to obtain a p-value for a factor covariate 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 covariate ($x_i \in \{0,1\}$).

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 $\texttt{log(Hg}_{\texttt{urin}}\texttt{)}$ depending on the age of the participants:

```
## [1] "Missing HG_URING.csv"
```

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

→ Children and mothers 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 ($\text{age}_i \cdot \text{mother}_i$):

```
## [1] "Missing HG DATA"
```

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?

→ Clicker exercise http://www.klicker.uzh.ch/bkx

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

```
print('MISSING HG DATA')
```

```
## [1] "MISSING HG DATA"
```

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

[1] “missing HG DATA”

(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):

```
## [1] "missing HG DATA"
```

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 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!

Multiple vs. many single regressions

Question: Given multiple regression covariates $x^{(1)}, x^{(2)}, \ldots$. 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 covariates $x^{(1)}, x^{(2)}, \ldots$. 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):

Why?

Illustration

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

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

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

Thus the correct 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):

> *[...] 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

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):

```
## [1] "missing HG data"
```

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 obervations.
- Plot residuals against the covariates.

```
## [1] "Missing HG data"
```

Again, no pattern = good.

QQ-plot

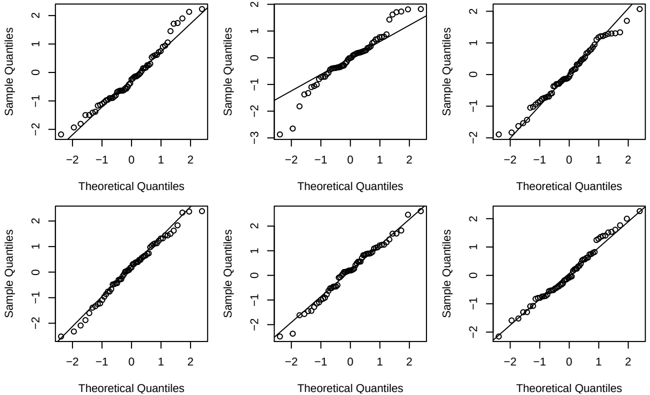
The outlier recorded above is also visible in the (well-known) QQ-plot, which is useful to check for normal distribution of residuals (assumption c):

```
## [1] "Missing HG data"
```

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 $\epsilon_i \sim N(0,1)$ 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 (standardized) residuals $\sqrt{|R_i|}$ against the fitted values $\hat{y}_i$. There should be **no trend** (Again Hg example):

```
## [1] "missing HG data"
```

Leverages (“Hebel”)

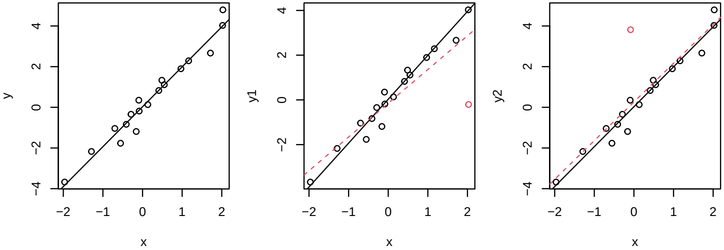
To understand the leverage plot, we need to introduce the idea of the *leverage* (“Hebel”).

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

http://www.klicker.uzh.ch/bkx

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.

→ 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):

```
## [1] "Missing HG data"
```

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 outcome or the covariables.
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?

```r
print('missing HG data')
```

```
## [1] "missing HG data"
```

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

```
## [1] "missing HG data"
```

Comparison to the model with log-transformed response:

```
## [1] "missing HG data"
```

```
## [1] "missing HG data"
```

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 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:

```
print('missing HG data')
```

```
## [1] "missing HG data"
```

```
#r4.urin.mother <- lm(log10(Hg_urin) ~  smoking + sqrt(amalgam) + sqrt(fis
```

```
## [1] "missing HG data"
```

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 the covariates and the response.
2. If not, ask whether the model has been misspecified. Do reasonable transformations of the response or the covariates 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!

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 40):

```
## [1] "missing HG data"
```

```
## [1] "missing HG data"
```

Feedback about today's lecture

Please give us your opinion about the lecture regarding

- unclearest ("muddiest") point
- the take-home message of today.

via this link:

http://www.klicker.uzh.ch/bkx

Thank you!!