Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Week 4: Multiple linear regression (finalize) / Residual analysis / Checking modelling assumptions

16./17. March 2017

Overview (todo: check)

- 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

- Chapter 3.3 in *Lineare Regression*
- To do

Recap of last week I

to do

Recap of last week I

Last week we introduced binary and factor covariates that allowed for group-specific intercepts.

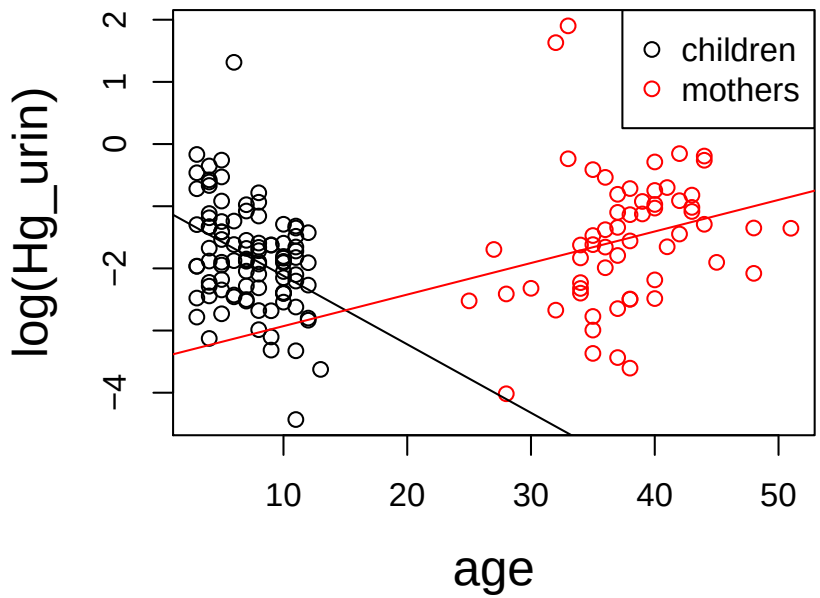
Group-specific slopes / Interactions

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

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

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

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).

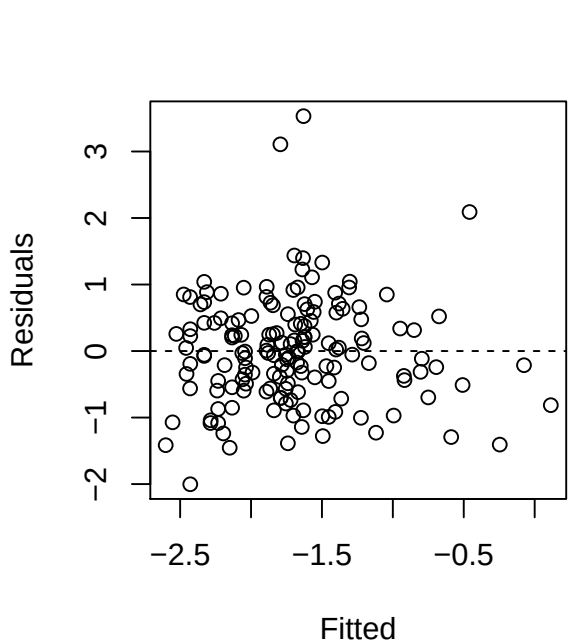
Remember (from last week), however, that the Hg model also included smoking status, amalgam fillings and fish consumption as important predictors. It is very straightforward to just include these predictors in model (1), which leads to the following model

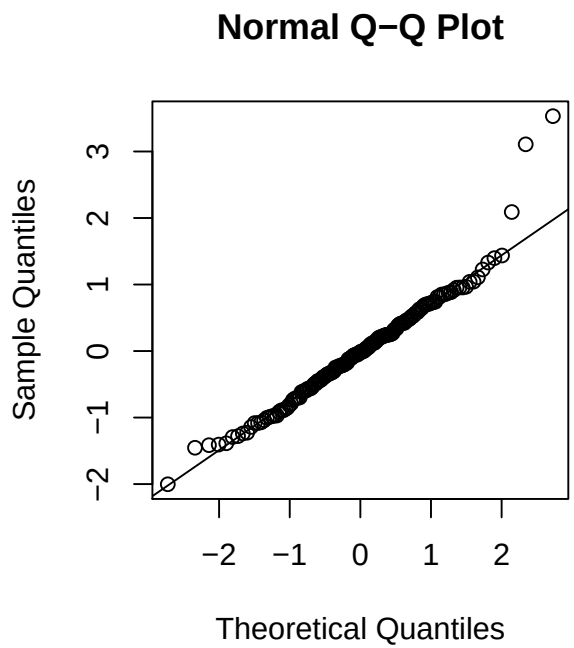
```
> r.hg <- lm(log(Hg_urin)~  mother * age + smoking + amalgam + fish,d.hg)
```

	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:





Multiple vs. many single regressions

Question: I find group-specific intercepts and interactions too complicated. Could I simply fit separate models for each group?

Multiple vs. many single regressions

Question: I find group-specific intercepts and interactions too complicated. Could I simply fit separate models for each group?

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?

Chapter 3.3c in the Stahel script illustrates the point on four artificial examples. The model is 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 (0/1)