

Lecture 5: Binary/categorical explanatory variables, and interactions

BIO144 Data Analysis in Biology

Stephanie Muff, Owen Petchey & Uriah Daugaard

University of Zurich

~~15~~ March, 2024

18

- ▶ Binary and categorical explanatory variables
- ▶ Interactions between explanatory variables
- ▶ Multiple vs. many single regressions

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

- ▶ Interpretation 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 $y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_m x_i^{(m)} + \epsilon_i$.

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**.

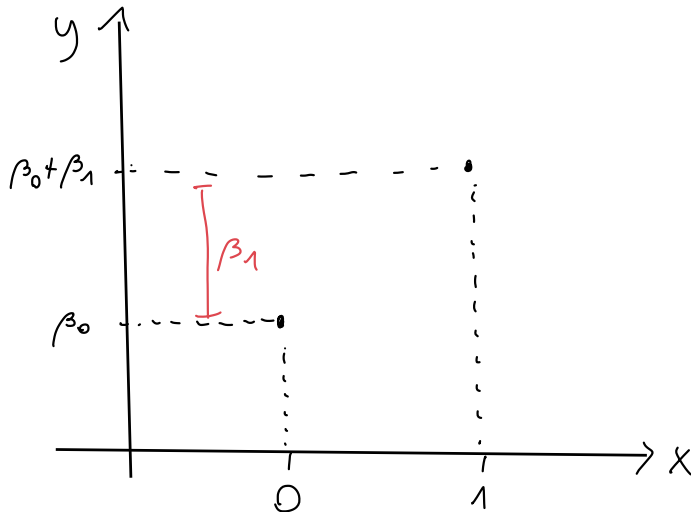
One very frequent data type are **binary** variables, that is, variables that can only attain values 0 or 1.

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:

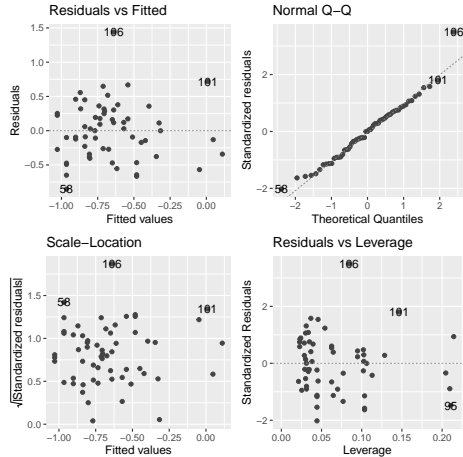
$$\log(Hg_{urin})_i = \beta_0 + \beta_1 \cdot x_i^{(1)} + \beta_2 \cdot x_i^{(2)} + \beta_3 \cdot x_i^{(3)} + \epsilon_i ,$$

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	<i>p</i> -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.

$$\text{Smokers: } y_i = -1.03 + 0.26 + 0.098 \cdot \text{amalgam}_i + 0.032 \cdot \text{fish}_i + \epsilon_i$$

$$\text{Non-smokers: } y_i = -1.03 + 0.098 \cdot \text{amalgam}_i + 0.032 \cdot \text{fish}_i + \epsilon_i$$

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 $x^{(1)}, \dots, x^{(k)}$ can then be included as a binary variable in the model

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

However: this model is **not identifiable**. We could add a constant to $\beta_1, \beta_2, \dots, \beta_k$ 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)

$n = 6$

Species	$x^{(A)}$	$x^{(B)}$	$x^{(C)}$
A	1	0	0
B	0	1	0
C	0	0	1
A	1	0	0
A	1	0	0
C	0	0	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 $x^{(1)}, \dots, x^{(k)}$ can then be included as a binary variable in the model

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

However: this model is **not identifiable**. We could add a constant to $\beta_1, \beta_2, \dots, \beta_k$ 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 $\beta_1 = 0$)

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

Sketch (2)

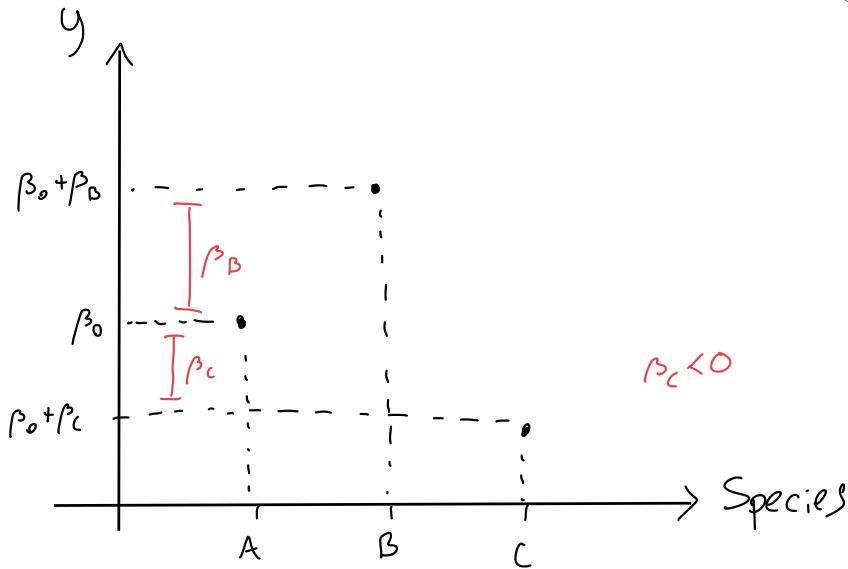
$n=6$

Species	$x^{(A)}$	$x^{(B)}$	$x^{(C)}$	Response y
A	1	0	0	β_0
B	0	1	0	$\beta_0 + \beta_B$
C	0	0	1	$\beta_0 + \beta_C$
A	1	0	0	β_0
A	1	0	0	β_0
C	0	0	1	$\beta_0 + \beta_C$

reference

$$y = \beta_0 + \cancel{\beta_A x^{(A)}} + \beta_B x^{(B)} + \beta_C x^{(C)} + \epsilon_i$$

Sketch (3)



!!Important to remember!!

(Common aspect that leads to confusion!)

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

→ 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
→ it has less “freedom”
- ▶ For example, to calculate $SSQ^{(Y)} = \sum_{i=1}^n (y_i - \bar{y})^2$ we need \bar{y}
 - ▶ Now let's say we know that $\bar{y} = 4$ & $n = 5$, 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 \bar{y} & n , we know that the fifth value must be $\square = 4 \Rightarrow$ the fifth value is no longer free to vary
 - ▶ Thus, for the calculation of $SSQ^{(Y)}$ we have $n - 1$ degrees of freedom (DF), because we used 1 DF to calculate \bar{y} from y_i (in other words, \bar{y} and y_i are not independent)
- ▶ The DF are the number of data points that are free to vary, given what we want to calculate
 - ▶ **Note:** We lose one DF for every beta parameter that we want to estimate!

Sketch (about degrees of freedom, 2)

Species	y	DF = n - 2 n = 8
A	2	
A	4	$\bar{y}_A = 4$
A	6	
A	<input type="checkbox"/> - not free	
<hr/>		
B	7	$\bar{y}_B = 6$
B	5	
B	6	
B	<input type="checkbox"/> - not free	

→ f: f a regression

y ~ species

We estimate

$$y = \beta_0 + \beta_1 \cdot \text{species} + \varepsilon$$

(2) beta parameters

$$\Rightarrow \text{DF} = n - (2)$$

Example: Earthworm study

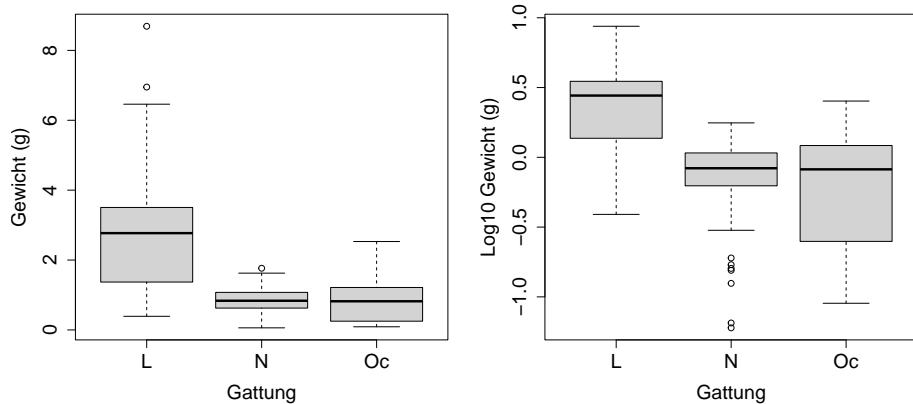
(Angelika von Förster und Burgi Liebst)

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 berechenbar, 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 genera (Lumbricus, Nicodrilus, Octolasion), weight, stomach circumference (Magenumfang).

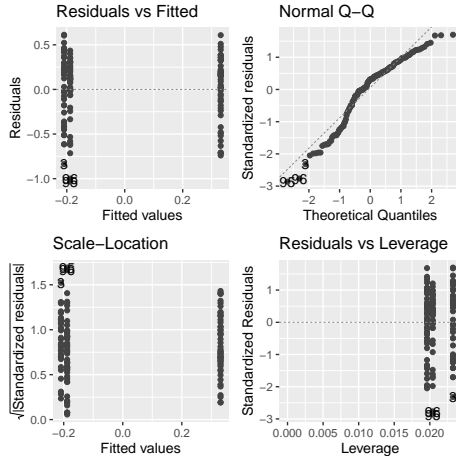
Data inspection suggests that the three genera have different weights:



Formulate a model with $\log_{10}(\text{Gewicht})$ 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

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

- ▶ 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?

$$DF = 143 - \overbrace{(3-1)}^{\text{Gattung}} - \overbrace{1}^{\text{intercept}} = 140$$

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 genus:

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

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

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

$\beta_2 =$ difference in intercept for Gattung N

$\beta_3 =$ difference in intercept for Gattung Oc

Interpreting the results II

Question: Is the “Gattung” explanatory variable relevant in the model, that is, do the model intercepts differ for the three genera?

Problem: The p -values of the t-test for each of the worm genus 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 zum Vergleich von Modellen. Die Frage sei, ob die q Koeffizienten $\beta_{j_1}, \beta_{j_2}, \dots, \beta_{j_q}$ 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:

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

$\text{SSQ}^{(E)*}$ 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 Nullhypothese: $T \sim \mathcal{F}_{q, n-p}$, 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: does the categorical variable *Gattung* explain a significant amount of variability?

$$F\text{-statistic} = \frac{\text{additional variability explained by including } Gattung}{\text{residual variability not explained}} = \frac{\overbrace{(SSE^* - SSE)}^{\geq 0} / q}{SSE / (n-p)} \sim F\text{-distr. with } q \text{ \& } n-p \text{ degrees of freedom}$$

$$SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

$q = \text{number of categories} - 1 = k - 1$

$n = \text{number of data points}$

$p = \text{number of beta parameters}$

$H_0: \beta_N = \beta_{0c} = 0$

$H_A: \text{at least one is } \neq 0$

$SSE^* = SSE \text{ of the model with that the categorical variable}$

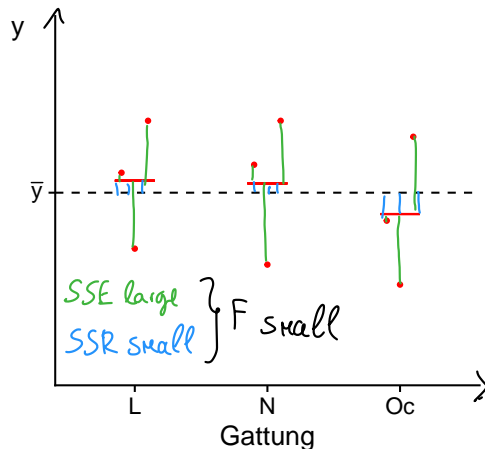
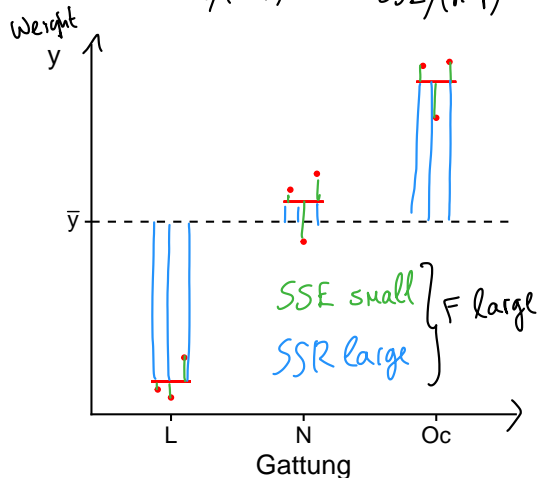
Remember, $SST = SSR + SSE$

Special case: if the categorical variable is the only variable in the model, then $SSR^* = 0 \Rightarrow SSE^* = SST$

$$F\text{-stat} = \frac{(SST - SSE) / q}{SSE / (n-p)} = \frac{SSR / q}{SSE / (n-p)} \sim F_{q, n-p} ; SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2 ; SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

F-test (sketch)

$$F\text{-stat} = \frac{(SST - SSE)/q}{SSE/(n-p)} = \frac{SSR/q}{SSE/(n-p)} \sim F_{q, n-p} ; \quad \boxed{SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2} ; \quad \boxed{SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$$



F-test for the earthworms

ANalysis Of VAriance

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
```

$$\frac{9.2}{2} = MS_{\text{Gattung}} = 4.6$$

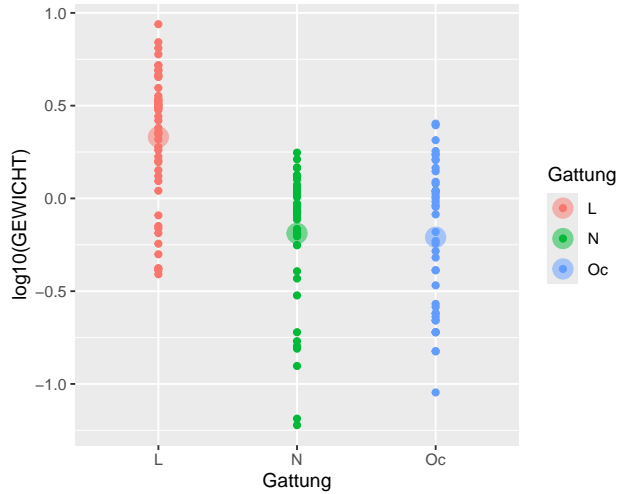
$$\frac{18.64}{140} = 0.13 = MS_{\text{Errors}}$$

$$F = \frac{MS_{\text{Gattung}}}{MS_{\text{Errors}}} = 34.6$$

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 genera ("Gattung is relevant").

Plotting the earthworms results

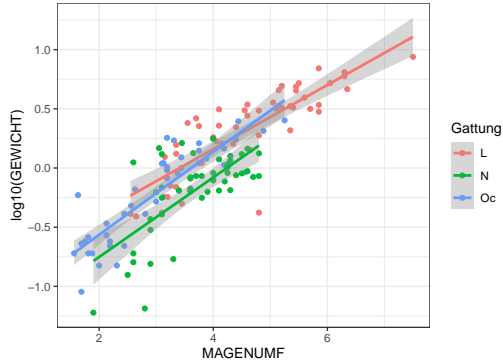


That was a lot. . .

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.

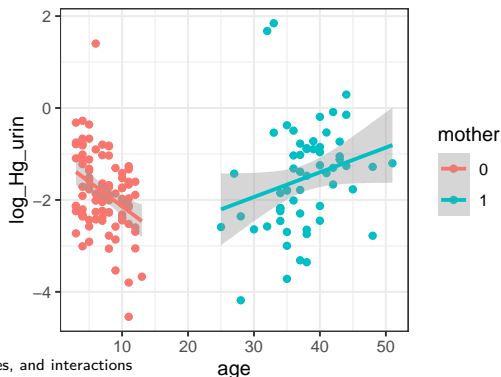
Homework

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 $\log(\text{Hg}_{\text{urin}})$ 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 ($x_i = 1$): $\hat{y}_i = \underbrace{\beta_0 + \beta_1}_{\text{Intercept}} + \underbrace{(\beta_2 + \beta_3)}_{\text{slope}} \text{age}_i$

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

Difference in slopes

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)
β_0	## (Intercept)	-1.0685772	0.25997341	-4.110333	6.442221e-05
β_1	## mother1	-2.4811450	0.93896995	-2.642411	9.093456e-03
β_2	## age	-0.1071729	0.03321050	-3.227078	1.531867e-03
β_3	## mother1:age	0.1609570	0.04083104	3.942026	1.229208e-04

Interpretation:

Mothers: $\hat{y}_i = -1.06 + (-2.48) + \overbrace{(-0.10 + 0.16)}^{>0} \cdot \text{age}_i$

Children: $\hat{y}_i = -1.06 + \underbrace{(-0.10)}_{<0} \cdot \text{age}_i$

- ▶ 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?

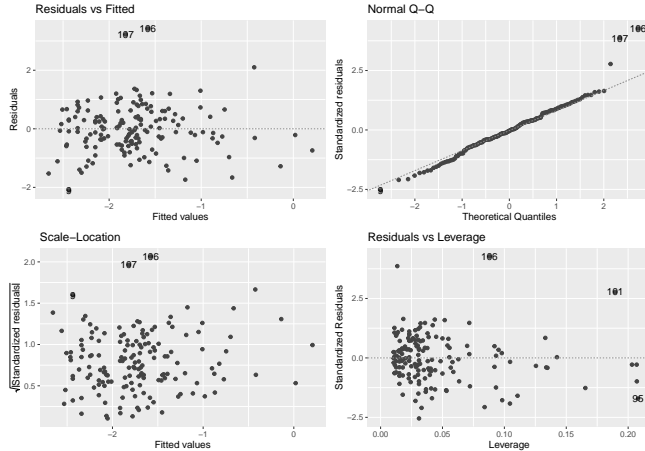
Remember, however, that the Hg model also included smoking status, amalgam fillings and fish consumption as important predictors. It is very straightforward to just include these predictors in model (1), which leads to the following model

```
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:

- ▶ \log
- ▶ $\sqrt{\cdot}$
- ▶ \sin, \cos, \dots

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

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

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

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**.

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

Thus you will **always** need the F -test `anova()` to obtain a p -value for a categorical 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).