Kurs Bio144: Datenanalyse in der Biologie

Lecture 7: ANCOVA, short introduction to Linear Algebra

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

University of Zurich

18 December, 2020

Overview

- ANCOVA
- Introduction to linear Algebra

Note: ANCOVA = ANalysi of COVAriance (Kovarianzanalyse)

Course material covered today

- "Getting Started with R" chapter 6.3
- "Lineare regression" chapters 3.A (p. 43-45) and 3.4, 3.5 (p. 39-42)

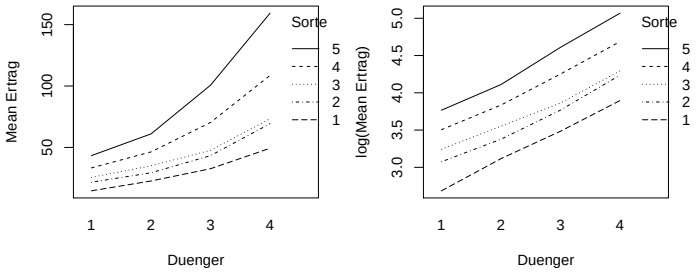
Recap of ANOVA

- ANOVA is a method to test if the means of two or more groups are different.
- Post-hoc tests and contrasts, including correction for p-values, to understand the differences between the groups.
- Two-way ANOVA for factorial designs, interactions.
- ANOVA is a special case of linear regression with categorical covariates.

Recap of two-way ANOVA example

Remember: Influence of four levels of fertilizer (DUENGER) on the yield (ERTRAG) on 5 species (SORTE) of crops was investigated. For each DUENGER $\times$ ERTRAG combination, 3 repeats were taken.

Interaction plot with ERTRAG and log(ERTRAG) as response:



Remember: We used log(ERTRAG), because the residual plots were otherwise not ok.

```
r.duenger2 <- lm(log(ERTRAG) ~ DUENGER*SORTE,d.duenger)
anova(r.duenger2)
```

```
## Analysis of Variance Table
##
## Response: log(ERTRAG)
##               Df  Sum Sq Mean Sq  F value Pr(>F)
## DUENGER        3 11.6917  3.8972 854.0505 <2e-16 ***
## SORTE          4  8.5202  2.1300 466.7851 <2e-16 ***
## DUENGER:SORTE 12  0.0929  0.0077   1.6958 0.1045
## Residuals     40  0.1825  0.0046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Questions:

- Number of parameters?
- Degrees of freedom (60 data points)?
- Interpretation?

```
summary(fit.duenger2)
```

```
##
## Call:
## lm(formula = log(ERTRAG) ~ DUENGER * SORTE, data = d.duenger)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.120968 -0.045595  0.008984  0.049072  0.102175
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.68505    0.03900  68.846  < 2e-16 ***
## DUENGER2            0.43165    0.05516   7.826 1.36e-09 ***
## DUENGER3            0.79997    0.05516  14.504  < 2e-16 ***
## DUENGER4            1.21152    0.05516  21.966  < 2e-16 ***
## SORTE2              0.38979    0.05516   7.067 1.51e-08 ***
## SORTE3              0.55799    0.05516  10.117 1.38e-12 ***
## SORTE4              0.82018    0.05516  14.870  < 2e-16 ***
## SORTE5              1.08169    0.05516  19.612  < 2e-16 ***
## DUENGER2:SORTE2    -0.12949    0.07800  -1.660    0.105
## DUENGER3:SORTE2    -0.10613    0.07800  -1.361    0.181
## DUENGER4:SORTE2    -0.04924    0.07800  -0.631    0.531
## DUENGER2:SORTE3    -0.12180    0.07800  -1.562    0.126
## DUENGER3:SORTE3    -0.18034    0.07800  -2.312    0.026 *
## DUENGER4:SORTE3    -0.16061    0.07800  -2.059    0.046 *
## DUENGER2:SORTE4    -0.10138    0.07800  -1.300    0.201
## DUENGER3:SORTE4    -0.05311    0.07800  -0.681    0.500
## DUENGER4:SORTE4    -0.02954    0.07800  -0.379    0.707
## DUENGER2:SORTE5    -0.08779    0.07800  -1.125    0.267
## DUENGER3:SORTE5     0.04370    0.07800   0.560    0.578
## DUENGER4:SORTE5     0.09014    0.07800   1.156    0.255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06755 on 40 degrees of freedom
## Multiple R-squared:  0.9911,  Adjusted R-squared:  0.9869
```

Analysis of Covariance (ANCOVA)

An ANCOVA is an analysis of variance (ANOVA), including also at least one continuous covariate.

ANCOVA unifies several concepts that we approached in this course so far:

- Linear regression
- Categorical covariates
- Interactions (of continuous and categorical covariates)
- Analysis of Variance (ANOVA)

As such, it is a **special case of the linear regression model**.

Given a categorical covariate x_i and a continuous covariate z_i. Then the ANCOVA equation (without interactions) is given as

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

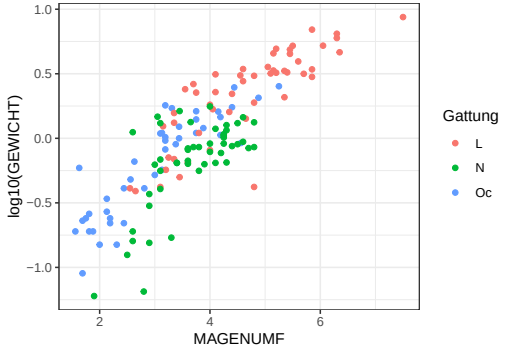
where $x_i^{(k)}$ is the kth dummy variable ($x_i^{(k)}$=1 if ith observation belongs to category k, 0 otherwise).

Note 1: It is straightforward to add an interaction of x_i with z_i.

Note 2: Again, for identifiability reason, we typically set $\beta_1 = 0$.

Once more: the earthworms

"Magenumfang" was used to predict "Gewicht" of the worm, including as covariate also the worm species.



Categorical and continuous covariates were used to predict a continuous outcome → ANCOVA.

```
r.lm <- lm(log(GEWICHT) ~  MAGENUMF + Gattung,d.wurm)
summary(r.lm)$coef
```

```
##               Estimate Std. Error     t value     Pr(>|t|)
## (Intercept) -2.5355459 0.22147279 -11.4485663 8.617670e-22
## MAGENUMF     0.7118725 0.04528843  15.7186392 1.232126e-32
## GattungN    -0.5151344 0.11009219  -4.6791186 6.760621e-06
## GattungOc   -0.0907298 0.12791000  -0.7093254 4.793107e-01
```

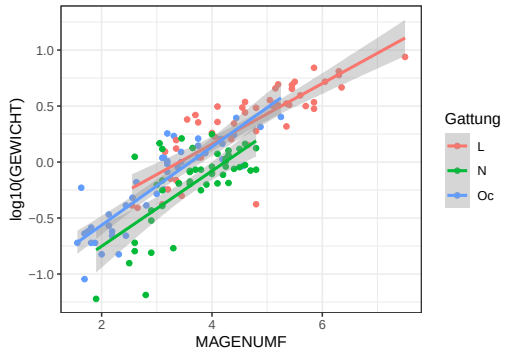
Important: The p-values for the entries `GattungN` and `GattungOc` are not very meaningful (why?).

To understand if "Gattung" has an effect, **we need to carry out an F-test** $\rightarrow$ ANOVA table:

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

```
## Analysis of Variance Table
##
## Response: log(GEWICHT)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## MAGENUMF    1 104.866 104.866  409.69 < 2.2e-16 ***
## Gattung     2   7.177   3.589   14.02 2.842e-06 ***
## Residuals 139  35.579   0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We also included an **interaction term** between `MAGENUMF` and `Gattung` to allow for different slopes:



→ We again need the F-**test** to check whether the respective interaction term is needed:

```
r.lm2<- lm(log(GEWICHT) ~  MAGENUMF * Gattung,d.wurm)
anova(r.lm2)
```

```
## Analysis of Variance Table
##
## Response: log(GEWICHT)
##                   Df  Sum Sq Mean Sq  F value    Pr(>F)
## MAGENUMF           1 104.866 104.866 414.4743 < 2.2e-16 ***
## Gattung            2   7.177   3.589  14.1835 2.521e-06 ***
## MAGENUMF:Gattung   2   0.917   0.458   1.8112    0.1673
## Residuals        137  34.662   0.253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

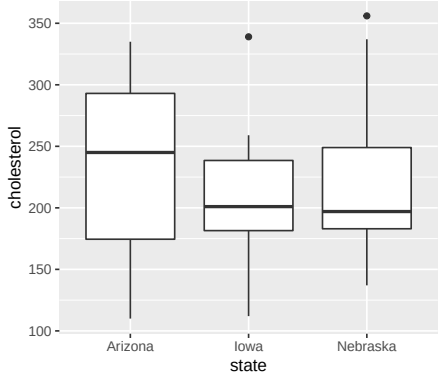
→ $p = 0.167$, thus interaction is probably not relevant.

A new example: cholesterol levels

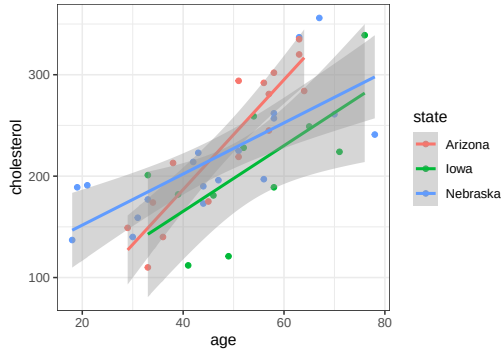
Example: Cholesterol levels [mg/ml] for 45 women from three US states (Iowa, Nebraska, Arizona), were measured.

Question: Do these levels differ between the states?

Age (years) may be a relevant covariable.



The scatter plot gives an idea about the model that might be useful here:



$\rightarrow$ We include state, age and the interaction of the two.

Doing the analysis:

```
r.lm <- lm(cholesterol ~ age*state,data=d.chol)
anova(r.lm)
```

```
## Analysis of Variance Table
##
## Response: cholesterol
##           Df Sum Sq Mean Sq F value    Pr(>F)
## age        1  96524   96524 61.8961 1.424e-09 ***
## state      2  11474    5737  3.6789   0.03438 *
## age:state  2  12665    6332  4.0606   0.02501 *
## Residuals 39  60819    1559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation?

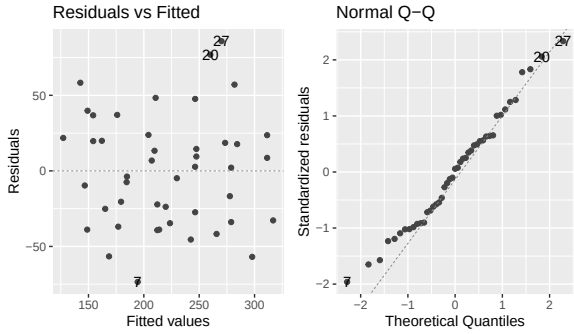
Compare the results from the previous slide to the estimated coefficients:

```
r.lm <- lm(cholesterol ~ age*state,data=d.chol)
summary(r.lm)$coef
```

```
##                       Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)       -29.895169 43.7353712 -0.6835467 4.983027e-01
## age                 5.416908  0.8679635  6.2409400 2.396876e-07
## stateIowa          65.706383 66.7677031  0.9841043 3.311303e-01
## stateNebraska     131.192935 50.8573164  2.5796276 1.377434e-02
## age:stateIowa      -2.178763  1.2672928 -1.7192264 9.350204e-02
## age:stateNebraska  -2.896470  1.0166558 -2.8490174 6.967607e-03
```

Note: The p-values for the age coefficient is not the same as in the ANOVA table.
Reason: `anova()` tests the models against one another in the **order** specified.

As always, some model checking is necessary:



$\rightarrow$ This seems ok.

An introduction to linear Algebra

Who has some knowledge of linear Algebra?

Overview

- The basics about
  - vectors
  - matrices
  - matrix algebra
  - matrix multiplication
- Why is linear Algebra useful?
- What does it have to do with data analysis and statistics?
- Regression equations in matrix notation.

Motivation

Why are vectors, matrices and their algebraic rules useful?

xample 1: The observations for a covariate x or the response y for all individuals $1 \leq i \leq n$ can be stored in a vector (vectors and matrices are always given in **bold** letters):

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}.$$

xample 2: Covariance matrices for multiple variables. Say we have $x^{(1)}$ and $x^{(2)}$. The covariance matrix is then given as

$$\begin{pmatrix} Var(x^{(1)}) & Cov(x^{(1)}, x^{(2)}) \\ Cov(x^{(1)}, x^{(2)}) & Var(x^{(2)}) \end{pmatrix}.$$