

# Lecture 6: ANOVA

## BIO144 Data Analysis in Biology

Stephanie Muff & Owen Petchey

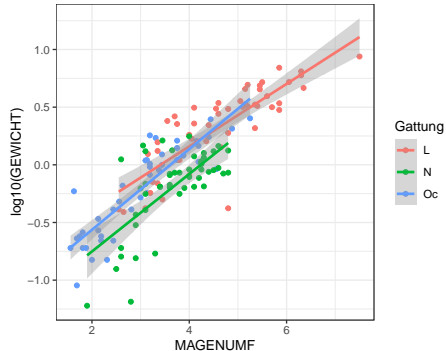
University of Zurich

25 March, 2021

## Recap of muddiest point from last week

Main topic: Fitting and interpreting models with interactions.

Let's go back to the earthworm example, and fit a model that allows species-specific intercepts and slopes:



```
r.lm <- lm(log10(GEWICHT) ~ MAGENUMF * Gattung,d.wurm)
summary(r.lm)
```

```
##
## Call:
## lm(formula = log10(GEWICHT) ~ MAGENUMF * Gattung, data = d.wurm)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.75318	-0.12834	0.01742	0.12268	0.59732

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.92394	0.13402	-6.894	1.82e-10 ***
MAGENUMF	0.27091	0.02816	9.620	< 2e-16 ***
GattungN	-0.49990	0.21454	-2.330	0.0213 *
GattungOc	-0.33921	0.17228	-1.969	0.0510 .
MAGENUMF:GattungN	0.06516	0.05289	1.232	0.2200
MAGENUMF:GattungOc	0.07894	0.04430	1.782	0.0769 .

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2185 on 137 degrees of freedom
## Multiple R-squared:  0.7652, Adjusted R-squared:  0.7566
## F-statistic: 89.29 on 5 and 137 DF,  p-value: < 2.2e-16
```

- ▶ Which are the interaction terms?
- ▶ Interpretation?

We have now actually fitted **three** models, one model for each species:

$$\text{L: } \hat{y}_i = -0.92 + -0.92 + 0.27 \cdot \text{MAGENUMF}$$

$$\text{N: } \hat{y}_i = -0.92 + -0.50 + 0.27 + 0.07 \cdot \text{MAGENUMF}$$

$$\text{O: } \hat{y}_i = -0.92 + -0.34 + 0.27 + 0.08 \cdot \text{MAGENUMF}$$

To remember:

- ▶ The “*Gattung*” terms in the model output are the **differences in intercepts** with respect to the reference category.
- ▶ The “*MAGENUMF:Gattung*” terms in the model output are the **differences in slopes** with respect to the reference category.

## Testing for an interaction term

If we want to find out if the interaction term for a categorical explanatory variable with more than two categories is relevant, we again need an  $F$ -test, that is, use the `anova()` function:

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

```
## Analysis of Variance Table
##
## Response: log10(GEWICHT)
##              Df    Sum Sq Mean Sq  F value    Pr(>F)
## MAGENUMF      1 19.7790 19.7790 414.4743 < 2.2e-16 ***
## Gattung       2  1.3537  0.6768  14.1835 2.521e-06 ***
## MAGENUMF:Gattung 2  0.1729  0.0864   1.8112  0.1673
## Residuals    137  6.5377  0.0477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here,  $p = 0.167$ , thus there is not much evidence that the three species differ in their regression slopes.

# Overview for today

- ▶ One-way ANOVA
- ▶ Post-hoc tests and contrasts
- ▶ Two-way ANOVA
- ▶ ANOVA as special cases of a linear model

Note:

ANOVA = ANalysis Of VAriance (Varianzanalyse)

# Course material covered today

The lecture material of today is based on the following literature:

- ▶ Chapter 12 from Stahel book “Statistische Datenanalyse”
- ▶ “Getting Started with R” chapters 5.6 and 6.2

# ANOVA and ANCOVA

ANOVA = Varianzanalyse

ANCOVA = Kovarianzanalyse

Introduction by Sir R. A. Fisher (1890-1962). He worked at the agricultural research station in Rothamstead (England). AN(C)OVA are/were therefore traditionally used to analyze agricultural experiments.

Central question of AN(C)OVA:

Are the means of two or more groups different?

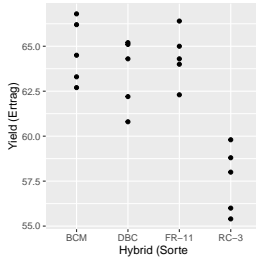


# Example: Yield of Hybrid-Mais breeds with increased resistance to “Pilzbrand”

(Source: W. Blanckenhorn, UZH)

Four different hybrid Mais breeds were grown to assess their yield. Each breed was grown at 5 different locations.

**Questions:** Are there differences in yield among the four hybrids?



We can test with ANOVA whether there are differences between the four breeds.

## One idea

To carry out pairwise  $t$ -tests between any two groups.

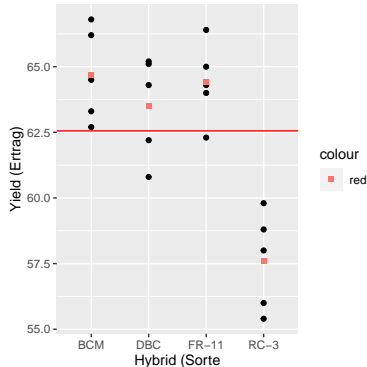
- ▶ How many tests would this imply?
- ▶ Why is this not a very clever idea?

## Better idea

Formulate a model that is able to **test simultaneously** whether there is an **overall difference between the groups**. That is, ask only **one question!**

This leads us to the

**Idea of the analysis of variance (ANOVA):** Compare the variability within groups ( $MS_E$ ) to the variability between the group means ( $MS_G$ ).





We formulate a model as follows:

$$y_{ij} = \mu_i + \epsilon_{ij} ,$$

where

- ▶  $y_{ij}$  = “Yield of the  $j^{\text{th}}$  plant of hybrid  $i$ ”
- ▶  $\mu_i$  = “Mean yield of hybrid  $i$ ”
- ▶  $\epsilon_{ij} \sim (0, \sigma^2)$  is an independent error term.

Typically, this is **rewritten as**

$$y_{ij} = \mu + \beta_i + \epsilon_{ij} ,$$

where  $\mu + \beta_i = \mu_i$  from above, thus the **group mean** of group  $i$ .

# One-way ANOVA (Einfaktorielle Varianzanalyse)

More generally, this leads us to the **One-way ANOVA**:

Assume we have  $g$  groups and in each group  $i$  there are  $n_i$  measurements of some variable of interest, denoted as  $y_{ij}$ . The model is then given as

$$\begin{aligned} y_{ij} &= \mu + \beta_i + \epsilon_{ij} \quad \text{for} \quad i = 1, \dots, g, \\ j &= 1, \dots, n_i, \\ \epsilon_{ij} &\sim (0, \sigma^2) \quad i.i.d. \end{aligned} \tag{1}$$

- ▶  $\mu$  plays the role of the **intercept**  $\beta_0$  in standard regression models.
- ▶ The estimation of  $\mu$ , and the  $\beta$  coefficients is again done by **least squares minimization**.
- ▶ The  $\epsilon_{ij} \sim (0, \sigma^2)$  *i.i.d.* assumption is again crucial, so **model checking** will be needed again.

Attention: Model (1) is overparameterized, thus an additional constraint is needed!  
Most popular:

- ▶  $\beta_1 = 0$  (**treatment contrast**; default in R). (such that  $\mu = \mu_1$ )

Interpretation: Group 1 is usually chosen such that it is some sort of **reference group** or **reference category**, for example a standard diet, while groups 2, 3, etc. correspond to novel diets whose effect is tested in an experiment.

- ▶  $\sum_i \beta_i = 0$  (**sum-to-zero contrast**).

Interpretation: The effects  $\beta_1, \beta_2$  etc give the deviation from the population averaged effect.

# ANOVA as a special case of a linear model

Model (1) is identical to the regression model with a categorical explanatory variable, see lecture 4.

**Interpretation: The categories are the different group memberships.**

Thus (assuming  $\beta_1 = 0$ ):

$$y_{ij} = \begin{cases} \mu + \epsilon_{ij}, & \text{for group 1} \\ \mu + \beta_2 + \epsilon_{ij}, & \text{for group 2} \\ \dots & \\ \mu + \beta_g + \epsilon_{ij}, & \text{for group } g . \end{cases}$$



# The ANOVA test: The $F$ -test

**Aim of ANOVA:** to test *globally* if the groups differ. That is:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_g \quad \text{or, equivalently} \quad \beta_2 = \dots = \beta_g = 0$$

$$H_1 : \text{At least two groups are different}$$

Remember lectures 4 and 5: We have already used the  $F$ -test for categorical variables (see  $F$ -test for the earthworms in lecture 4 or cooking rule on slide 6 in lecture 5).

This was equivalent to testing if all  $\beta$ s that belong to a categorical variable are  $=0$  at the same time.

→ Equivalent to testing if the categorical covariate is needed in the model.

This is **the very same problem here**, thus we need the  $F$ -test again!

## Variance decomposition

To derive the ingredients of the  $F$ -test, we look at the decomposition of variance (Remember this idea from lecture 4, slide 23):

$$\begin{aligned}
 \text{total variability} &= \text{explained variability} + \text{residual variability} \\
 SS_{total} &= SS_{\text{between groups}} + SS_{\text{within groups}} \\
 \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2 &= \sum_{i=1}^g n_i (\bar{y}_{\cdot i} - \bar{y})^2 + \sum_{i=1}^g \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{\cdot i})^2
 \end{aligned}$$

Degrees of freedom:

$$n - 1 = (g - 1) + (n - g)$$

From this:

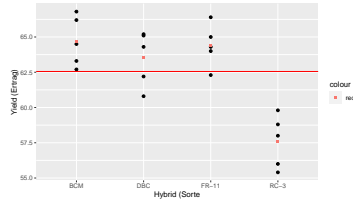
$$\left. \begin{aligned} MS_G &= \frac{SS_{\text{between}}}{g-1} \\ MS_E &= \frac{SS_{\text{within}}}{n-g} \end{aligned} \right\} \Rightarrow F = \frac{MS_G}{MS_E} \text{ is } \sim F_{g-1, n-g} \text{ distributed.}$$



# Interpretation of the $F$ statistic

- ▶  $MS_G$ : Quantifies the variability **between** groups.
- ▶  $MS_E$ : Quantifies the variability **within** groups.

## Example:



# Interpretation of the $F$ statistic II

- ▶  $F$  increases
  - ▶ when the group means become more different, or
  - ▶ when the variability within groups decreases.
- ▶ On the other hand,  $F$  decreases
  - ▶ when the group means become more similar, or
  - ▶ when the variability within groups increases.

→ The larger  $F$ , the less likely are the data seen under  $H_0$ .

▶ ANOVA App

[https://gallery.shinyapps.io/anova\\_shiny\\_rstudio/](https://gallery.shinyapps.io/anova_shiny_rstudio/)

# The ANOVA table

An overview of the results is typically given in an ANOVA table  
(Varianzanalysen-Tabelle):

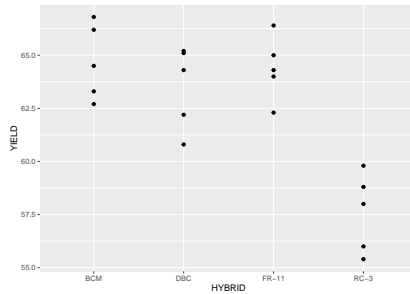
Variation	df	SS	MS = SS/df	F	p
Between groups	$g - 1$	$SS_G$	$MS_G$	$\frac{MS_G}{MS_E}$	$Pr(F_{g-1, n-g} >  F )$
Within groups	$n - g$	$SS_E$	$MS_E$		
Total	$n - 1$	$SS_{\text{total}}$			

# Our first ANOVA: Hybrid Mais example

HYBRID	LOCATION	YIELD
FR-11	NW	62
FR-11	NE	64
FR-11	C	64
FR-11	SE	65
FR-11	SW	66
BCM	NW	63
BCM	NE	63
BCM	C	66
BCM	SE	67
BCM	SW	64
DBC	NW	61
DBC	NE	64
DBC	C	65
DBC	SE	62
DBC	SW	65
RC-3	NW	55
RC-3	NE	56
RC-3	C	60
RC-3	SE	58
RC-3	SW	59

```
glimpse(d.mais)
```

```
## Rows: 20
## Columns: 3
## $ HYBRID   <chr> "FR-11", "FR-11", "FR-11", "FR-11", "FR-11", "BCM", "BCM", ...
## $ LOCATION <chr> "NW", "NE", "C", "SE", "SW", "NW", "NE", "C", "SE", "SW", ...
## $ YIELD    <dbl> 62.3, 64.0, 64.3, 65.0, 66.4, 63.3, 62.7, 66.2, 66.8, 64.5...
```



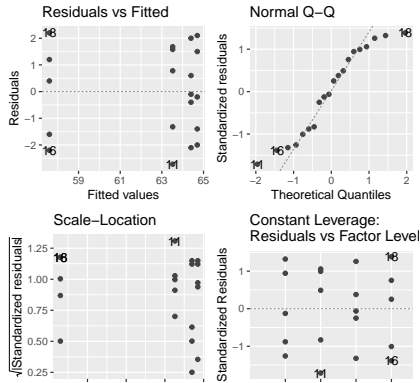


# Hybrid-Mais example – Estimation

Using the `lm()` function in R and then look at the ANOVA table:

```
r.mais <- lm(YIELD ~ HYBRID, d.mais)
```

Model checking is identical to all we did so far, because we are **still working with linear models!**



**Always** when we needed to do an  $F$ -test and when categorical covariates were involved, the `anova()` table is required:

```
anova(r.mais)

## Analysis of Variance Table
##
## Response: YIELD
##          Df Sum Sq Mean Sq F value    Pr(>F)
## HYBRID      3 167.441   55.814   17.681 2.474e-05 ***
## Residuals  16  50.508    3.157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

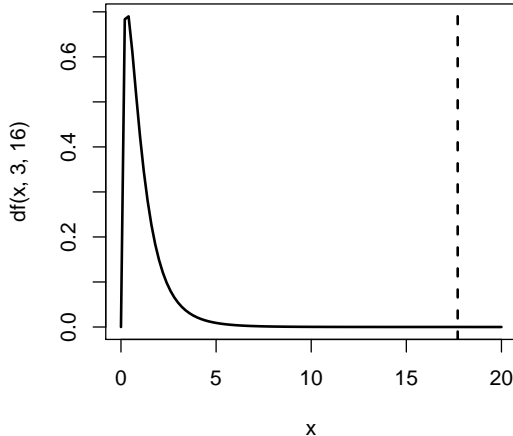
You can see that the value of  $F = 17.68$  is  $F$ -distributed with 3 and 16 degrees of freedom, and the  $p$ -value of the test “ $\beta_2 = \beta_3 = \beta_4 = 0$ ” is  $< 0.0001$ .

**Conclusion:** There are differences among the four groups!

→ This is equivalent to “The group variable is relevant for the model”.

**Exercise:** Look at the table a bit closer. How are Df, Sum Sq, Mean Sq, F value and  $\Pr(<F)$  related?

The  $F$ -distribution with 3 and 16 degrees of freedom, as well as the estimated value  $F=17.68$ :



# What happens if you apply `summary()` to the `lm()` object?

```
summary(r.mais)
```

```
##
## Call:
## lm(formula = YIELD ~ HYBRID, data = d.mais)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-2.72	-1.45	0.15	1.52	2.20

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	64.7000	0.7946	81.427	< 2e-16 ***
HYBRIDDBC	-1.1800	1.1237	-1.050	0.309
HYBRIDFR-11	-0.3000	1.1237	-0.267	0.793
HYBRIDRC-3	-7.1000	1.1237	-6.318	1.02e-05 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.777 on 16 degrees of freedom
## Multiple R-squared:  0.7683, Adjusted R-squared:  0.7248
## F-statistic: 17.68 on 3 and 16 DF,  p-value: 2.474e-05
```

The table contains the estimates of the intercept 64.70 ( $\mu$  in ANOVA notation,  $\beta_0$  in regression notation), and estimates for  $\beta_2$ ,  $\beta_3$ ,  $\beta_4$  (while the reference was set to  $\beta_1 = 0$ ).

## Post-hoc tests

**Still:** If the test  $\beta_2 = \dots = \beta_g = 0$  is rejected, a researcher is then often interested

1. in finding the actual group(s) that deviate(s) from the others.
2. in estimates of the pairwise differences.

Several methods to circumvent the problem of too many “significant” test results (type-I error) have been proposed. The most prominent ones are:

- ▶ Bonferroni correction
- ▶ Tukey **honest significant differences** (HSD) approach
- ▶ Fisher **least significant differences** (LSD) approach

## Bonferroni correction

**Idea:** If a total of  $m$  tests are carried out, simply divide the type-I error level  $\alpha_0$  (often 5%) such that

$$\alpha = \alpha_0 / m .$$

## Tukey HSD approach

**Idea:** Take into account the distribution of *ranges* (max-min) and design a new test.

## Fisher's LSD approach

**Idea:** Adjust the idea of a two-sample test, but use a larger variance (namely the pooled variance of all groups).

## Other contrasts

Sometimes additional comparisons are of interest. (Check also chapter 5.6.5 in GSWR)

## Choosing the reference category

Back to the Hybrid Mais example. R orders the categories alphabetically and takes the first level as reference category.

This can be changed manually:

```
levels(d.mais$HYBRID)
```

```
## NULL
```

```
d.mais <- mutate(d.mais, HYBRID = relevel(as.factor(HYBRID), ref="DBC"))
anova(lm(YIELD ~ HYBRID, d.mais))
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: YIELD
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## HYBRID      3 167.441   55.814   17.681 2.474e-05 ***
```

```
## Residuals 16  50.508    3.157
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(YIELD ~ HYBRID, d.mais))$coef
```

```
##           Estimate Std. Error   t value    Pr(>|t|)
```

```
## (Intercept)  63.52  0.7945754  79.9420714 2.974727e-22
```

```
## HYBRIDBCM     1.18  1.1236992  1.0501030 3.092739e-01
```

```
## HYBRIDFR-11   0.88  1.1236992  0.7831277 4.449899e-01
```

```
## HYBRIDRC-3   -5.92  1.1236992 -5.2683136 7.649526e-05
```



## Two-way ANOVA (Zweiweg-Varianzanalyse)

Example (from Hand et al. 1994 / Hothorn/Everitt "A Handbook of Statistical Analyses Using R"):

Experiment to study the weight gain of rats, depending on four diets. Protein amounts were either high or low, and the protein source was either beef or cereal. 10 rats for each diet were selected.

**Question:** How does diet affect weight gain?

**Complication:** This is a factorial design (gekreuzte Faktoren), because each combination of protein source (beef/cereal)  $\times$  category (high/low) is present ( $2 \times 2$  groups).

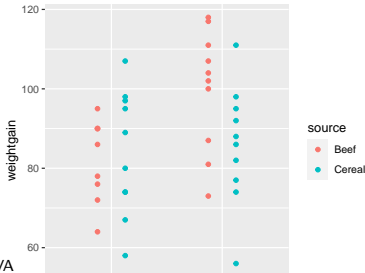
Design:

	beef	cereal
high	group 1	group 2
low	group 3	group 4

Start by looking at means and standard deviations in the groups, as well as at a graphical description of the means:

```
d.weightgain %>% group_by(source,amount) %>% summarise(meanW = mean(weightgain),sdW = sd(weightgain))
```

```
## # A tibble: 4 x 4
## # Groups:   source [2]
##   source amount meanW  sdW
##   <fct> <fct> <dbl> <dbl>
## 1 Beef   Low    79.2  13.9
## 2 Beef   High   100  15.1
## 3 Cereal Low    83.9  15.7
## 4 Cereal High    85.9  15.0
```



- \* Protein source (beef/cereal) seems less influential than the amount (high/low).
- \* Variances seem to be equal in the four groups.

## Two-way ANOVA – The model

In the presence of a **factorial design**, the idea is to add separate effects  $\beta_i$  (here  $i = 1, 2$ ) and  $\gamma_j$  (here  $j = 1, 2$ ) for the  $i$ th category of the first categorical variable and the  $j$ th category of the second explanatory variable:

Assume we have a factorial design with two categories  $\beta_i$  and  $\gamma_j$ , then the  $k$ th outcome in the category of  $i$  and  $j$ ,  $y_{ijk}$  is modelled as

$$y_{ijk} = \mu + \beta_i + \gamma_j + \epsilon_{ijk} \quad \text{with} \quad \epsilon_{ijk} \sim N(0, \sigma^2) \quad i.i.d.$$

Note: We again need additional constraints. Here we always use the R default ("**treatment contrasts**")

►  $\beta_1 = \gamma_1 = 0.$

Alternative:  $\sum_i \beta_i = \sum_j \gamma_j = 0$  (**sum-to-zero contrast**).

## Two way ANOVA

In R, a two-way ANOVA is as simple as one-way ANOVA, just add another variable:

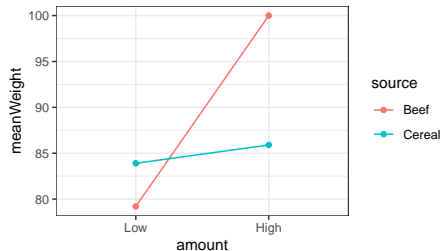
```
r.weight <- lm(weightgain ~ source + amount, d.weightgain)
anova(r.weight)
```

```
## Analysis of Variance Table
##
## Response: weightgain
##           Df Sum Sq Mean Sq F value    Pr(>F)
## source      1  220.9   220.90    0.9150  0.34501
## amount      1 1299.6  1299.60    5.3829  0.02596 *
## Residuals  37 8933.0   241.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretation: There seems to be a difference between low and high amounts of protein, but the source (beef/cereal) seems less relevant.

However: what if the additive model does not hold?

A so-called **interaction plot** helps to understand if the additive model is reasonable:



**Note:** if the additive model  $\beta_i + \gamma_j$  holds, the lines would be parallel.

However, these lines are **not parallel**, indicating that **there is an interaction** between amount and source!

In words: The amount (low/high amount treatment) has a different influence for the Beef and Cereal diets.

## Two-way ANOVA with interaction

- ▶ If the purely additive model is not correct, a more general model with an interaction term  $(\beta\gamma)_{ij}$  may be used:

$$y_{ijk} = \mu + \beta_i + \gamma_j + (\beta\gamma)_{ij} + \epsilon_{ijk} \quad \text{with} \quad \epsilon_{ijk} \sim N(0, \sigma^2) \quad i.i.d.$$

- ▶ As in linear regression, interactions allow for an **interplay between the variables**.
- ▶ In the rats experiment, increasing the amount from low to high has a different effect in the beef than in the cereal diet.
- ▶ Moreover: The plot on the previous slide shows that for the low amount of proteins case, the cereal diet leads to a larger average weight gain!

# Two-way ANOVA in R – Including an interaction

Let's include an interaction term in the rats example:

```
r.weight2 <- lm(weightgain ~ source * amount, d.weightgain)
anova(r.weight2)
```

```
## Analysis of Variance Table
##
## Response: weightgain
##           Df Sum Sq Mean Sq F value    Pr(>F)
## source      1  220.9   220.90   0.9879  0.32688
## amount      1 1299.6  1299.60   5.8123  0.02114 *
## source:amount 1   883.6   883.60   3.9518  0.05447 .
## Residuals   36 8049.4   223.59
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficient estimates can be obtained as follows:

```
summary(r.weight2)$coef
```

```
##           Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)      79.2    4.728577  16.7492235 1.416943e-18
## sourceCereal       4.7    6.687218   0.7028333 4.866800e-01
## amountHigh       20.8    6.687218   3.1104114 3.644273e-03
## sourceCereal:amountHigh -18.8    9.457155  -1.9879129 5.446757e-02
```

## Interpretation of the coefficients

This works in the same way as for categorical covariates in regression! To see this, let us estimate the means from the model. From the above output, we have [because of using treatment contrasts]:

$$\begin{aligned}\hat{\beta}_{beef} &= 0, \hat{\beta}_{cereal} = 4.7, \\ \hat{\gamma}_{low} &= 0, \hat{\gamma}_{high} = 20.8, \\ (\hat{\beta}\hat{\gamma})_{cereal/high} &= -18.8 \quad (\hat{\beta}\hat{\gamma})_{beef/high} = (\hat{\beta}\hat{\gamma})_{beef/low} = (\hat{\beta}\hat{\gamma})_{cereal/low} = 0.\end{aligned}$$

Therefore:

Group 1: beef / low	$\hat{y}_{beef,low} = 79.2 + 0 + 0 + 0 = 79.2$
Group 2: cereal / low	$\hat{y}_{cereal,low} = 79.2 + 4.7 + 0 + 0 = 83.9$
Group 3: beef / high	$\hat{y}_{beef,high} = 79.2 + 0 + 20.8 + 0 = 100$
Group 4: cereal / high	$\hat{y}_{cereal,high} = 79.2 + 4.7 + 20.8 - 18.8 = 85.9$



## A cautionary note

**Be careful:** In the presence of interactions, the  $p$ -values of the main effects can no longer be interpreted as before!

It is then required that separate “stratified” analyses are carried out. For example for “Beef” and “Cereal” protein sources:

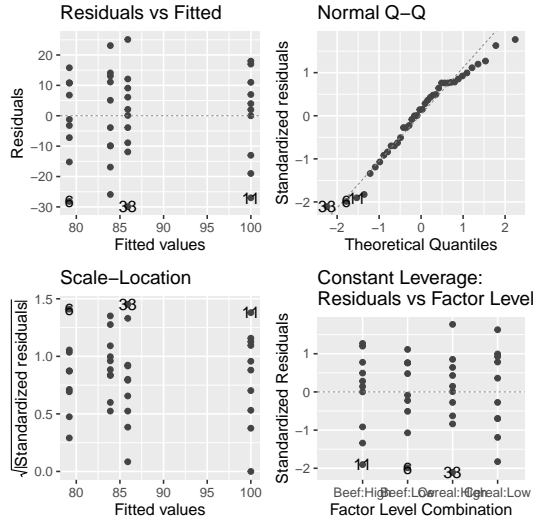
```
anova(lm(weightgain ~ amount,subset(d.weightgain,source=="Beef")))
```

```
## Analysis of Variance Table
##
## Response: weightgain
##          Df Sum Sq Mean Sq F value    Pr(>F)
## amount      1 2163.2  2163.20   10.253 0.00494 **
## Residuals  18 3797.6   210.98
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(weightgain ~ amount,subset(d.weightgain,source=="Cereal")))
```

```
## Analysis of Variance Table
##
## Response: weightgain
##          Df Sum Sq Mean Sq F value    Pr(>F)
## amount      1   20.0    20.00   0.0847 0.7744
## Residuals  18 4251.8   236.21
```

And finally, the model diagnostics:



## Exercise:

In an experiment the influence of four 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.

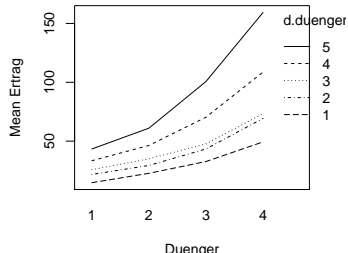
The data contain the following columns:

- ▶ DUENGER (4 levels)
- ▶ SORTE (5 levels)
- ▶ ERTRAG (continuous)

The first 10 rows of the data:

##	DUENGER	SORTE	ERTRAG
## 1	1	1	14
## 2	1	1	15
## 3	1	1	15
## 4	2	1	20
## 5	2	1	25
## 6	2	1	23
## 7	3	1	35
## 8	3	1	31
## 9	3	1	32
## 10	4	1	52

And the interaction plot:

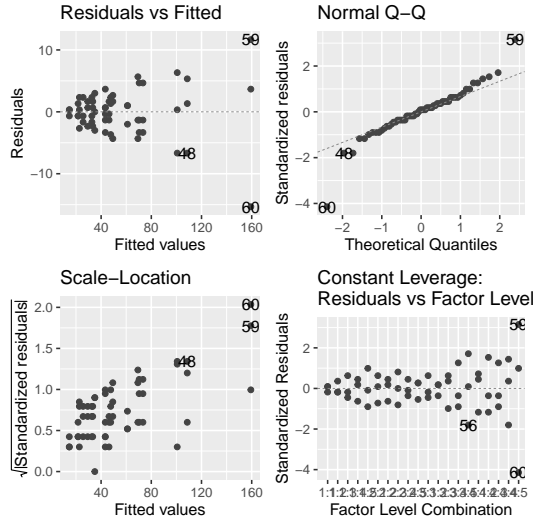


```
d.duenger <- mutate(d.duenger, SORTE=as.factor(SORTE), DUENGER=as.factor(DUENGER))  
r.duenger <- lm(ERTRAG ~ DUENGER*SORTE, d.duenger)  
#anova(r.duenger)
```

Look at the TA and the scale-location plots (next slide).

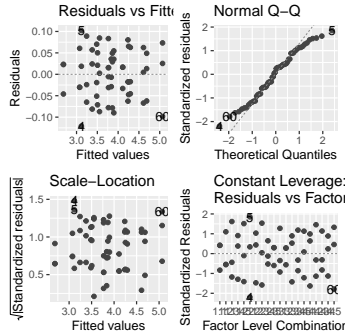
# What is the problem?

→ Interpretation? Ideas?



# Log-transform the response (ERTRAG) and repeat the analysis:

```
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		

Btw, the summary table with coefficients looks horrible and the *p*-values are not meaningful! (why?)

	Coefficient	95%-confidence interval	<i>p</i> -value
Intercept	2.69	from 2.61 to 2.76	< 0.0001
DUENGER2	0.43	from 0.32 to 0.54	< 0.0001
DUENGER3	0.80	from 0.69 to 0.91	< 0.0001
DUENGER4	1.21	from 1.10 to 1.32	< 0.0001
SORTE2	0.39	from 0.28 to 0.50	< 0.0001
SORTE3	0.56	from 0.45 to 0.67	< 0.0001
SORTE4	0.82	from 0.71 to 0.93	< 0.0001
SORTE5	1.08	from 0.97 to 1.19	< 0.0001
DUENGER2:SORTE2	-0.13	from -0.29 to 0.03	0.10
DUENGER3:SORTE2	-0.11	from -0.26 to 0.05	0.18
DUENGER4:SORTE2	-0.049	from -0.21 to 0.11	0.53
DUENGER2:SORTE3	-0.12	from -0.28 to 0.04	0.13
DUENGER3:SORTE3	-0.18	from -0.34 to -0.02	0.026
DUENGER4:SORTE3	-0.16	from -0.32 to -0.00	0.046
DUENGER2:SORTE4	-0.10	from -0.26 to 0.06	0.20
DUENGER3:SORTE4	-0.053	from -0.21 to 0.10	0.50
DUENGER4:SORTE4	-0.03	from -0.19 to 0.13	0.71
DUENGER2:SORTE5	-0.088	from -0.25 to 0.07	0.27
DUENGER3:SORTE5	0.044	from -0.11 to 0.20	0.58
DUENGER4:SORTE5	0.09	from -0.07 to 0.25	0.25

Questions: Number of parameters? Degrees of freedom (60 data points)?



## Some summary remarks

- ▶ The  $t$ -test to compare the mean of **two groups** is a **special case of ANOVA**.
- ▶ ANOVA is a **special case of the linear regression model**.
- ▶ ANOVA is often taught in separate lectures, although it could be integrated in a lecture on linear regression.
- ▶ ANOVA is traditionally most used to analyze **experimental data**, though this is changing...