Tools for complex data analysis

University of Koblenz-Landau 2018/19

Unity of the linear model and multiple linear regression

Ralf B. Schäfer

Learning targets

- Describe the mathematical basis for including categorical predictors
- Explain and interpret analysis of variance
- Explain and interpret linear models with multiple predictors
- Describe the issue of multiple inference

Learning targets

- Describe the mathematical basis for including categorical predictors
  - Outline the relation between the t-test or an ANOVA with one variable and a linear regression with a categorical predictor.
  - Interpret the regression coefficients for categorical predictors.
- Explain and interpret analysis of variance
  - For which research questions can you use ANOVA?
  - Describe the concept of sum of squares for ANOVA. How are they calculated?
  - What are the assumptions of ANOVA and how can you check them?

Learning targets

- Explain and interpret linear models with multiple predictors
  - Classify the types of linear models with multiple predictors.
  - Why does the F-statistic differ between linear regression and ANOVA output?
  - Explain the types of ANOVA and when they should be used.
  - What are main effects and interactions?
  - How do interactions complicate the interpretation of main effects?
  - Why should you standardize variables before multiple linear regression if interactions are present?
- Describe the issue of multiple inference
  - What is the issue of multiple inference?
  - When and how do you need to account for multiple inference?

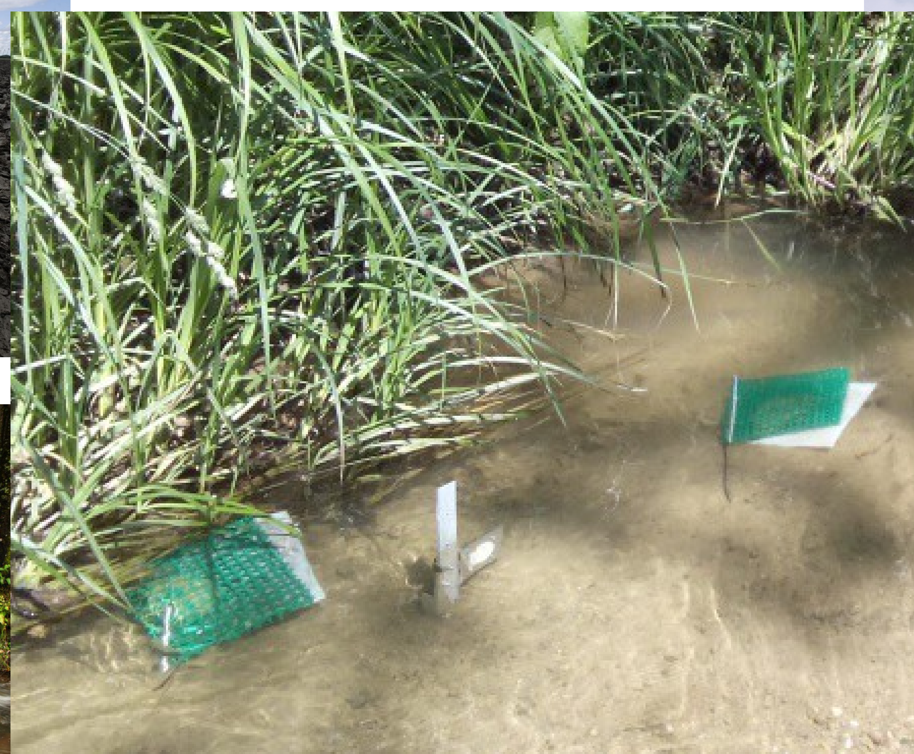
Unity of the linear model and multiple linear regression

Contents

Case study: Land use

<u>Research question:</u> Does land use influence the invertebrate leaf breakdown rate k_{inv} in streams?

<u>Study:</u> 29 streams with four major land uses in upstream catchment examined for leaf breakdown.

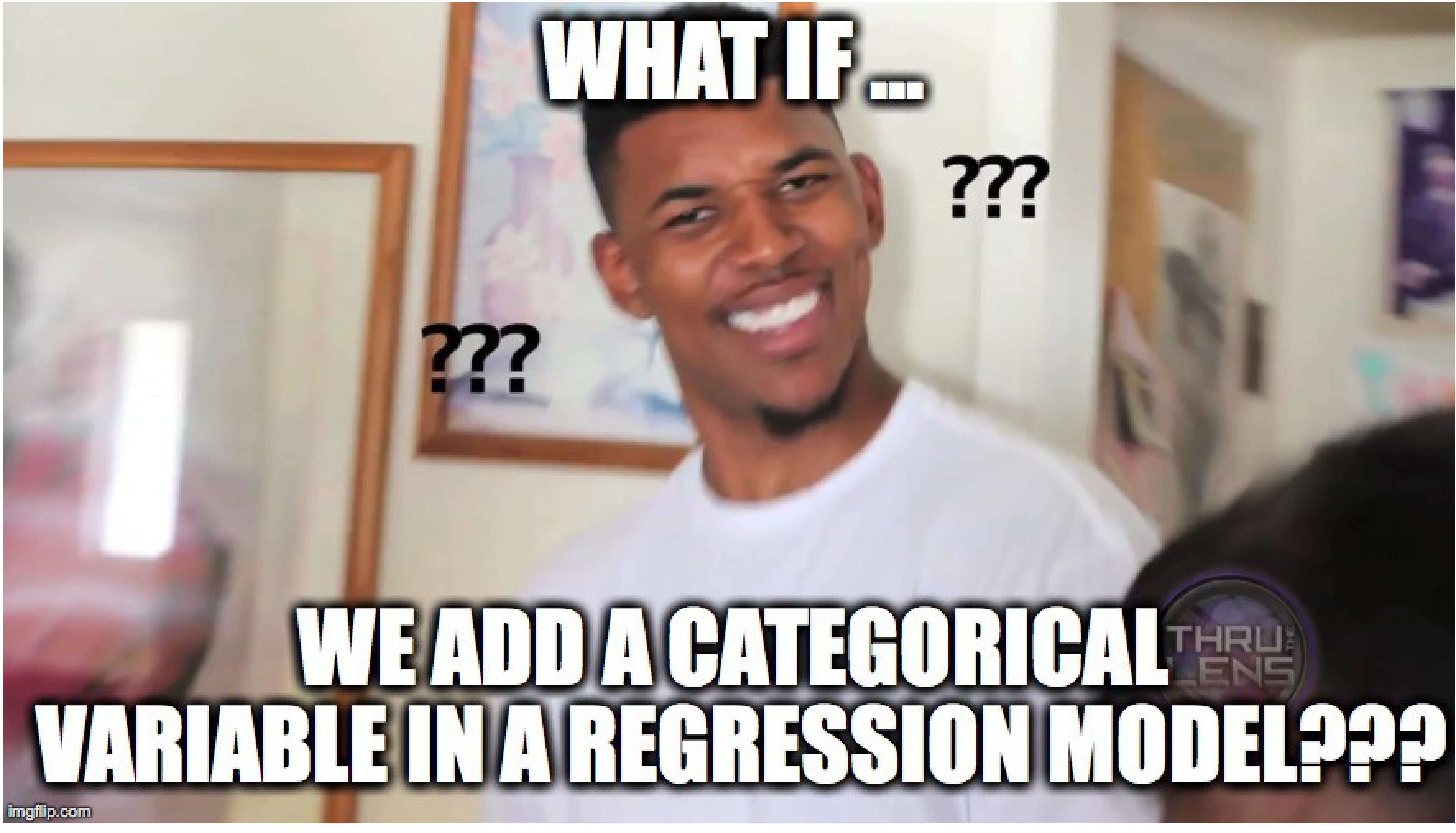


Pictures: Colourbox.de or own

Case study: Land use

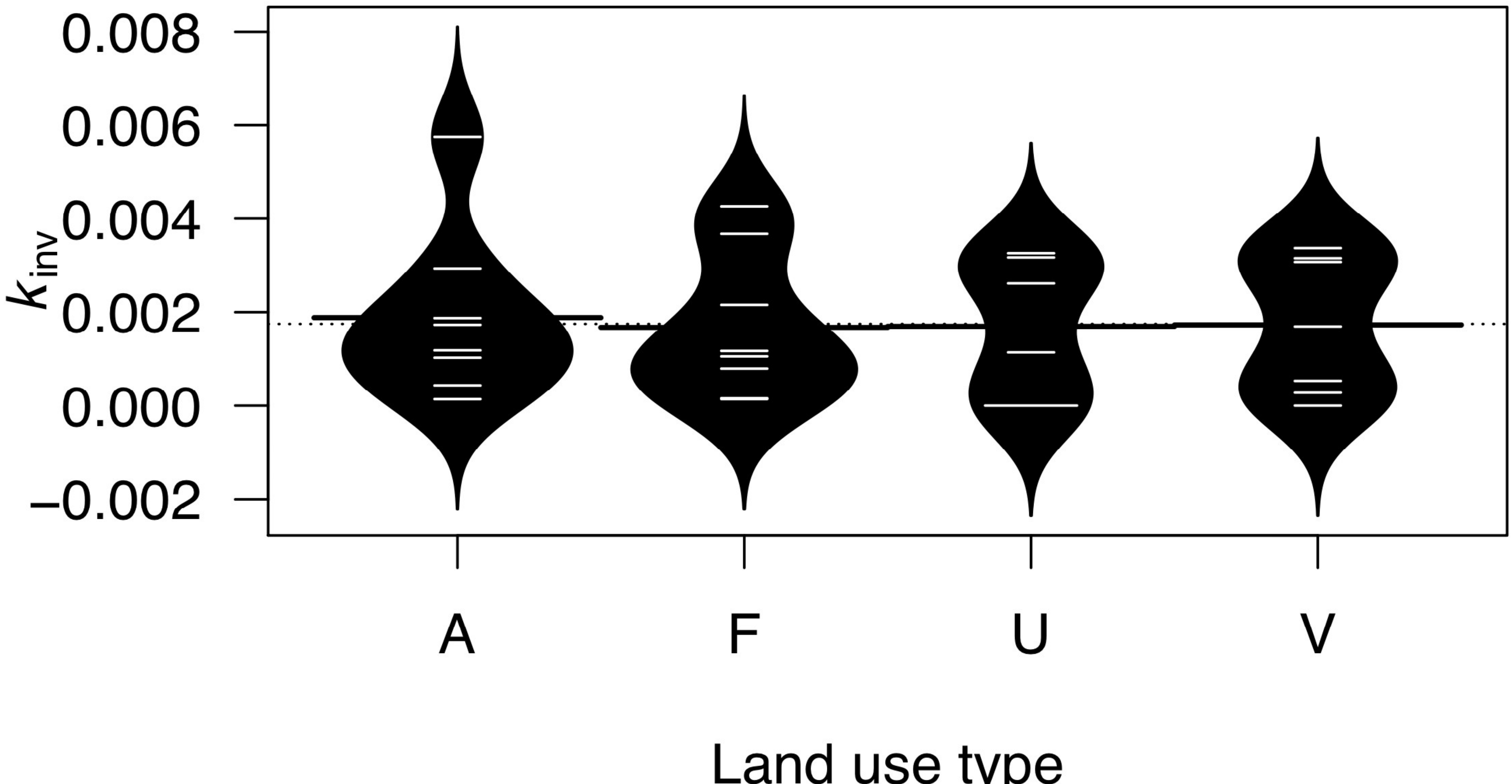
<u>Research question:</u> Does land use influence the invertebrate leaf breakdown rate k_{inv} in streams?

<u>Required analysis:</u> Relate k_{inv} (cont.) to land use type (categ.).



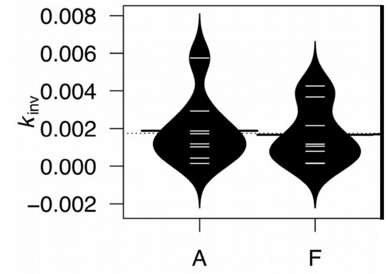
Case study: Land use

Graphical relationship of k_{inv} with land use type.



Categorical predictor in regression

Example with two-level predictor



Regression

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0018838  0.0005937   3.173  0.00677 **
## Land_typeF  -0.0002079  0.0008396  -0.248  0.80800
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.001679 on 14 degrees of freedom
## Multiple R-squared:  0.004362,    Adjusted R-squared:  -0.06676
## F-statistic: 0.06133 on 1 and 14 DF,  p-value: 0.808
```

t-test

```
##
##  Two Sample t-test
##
## data:  OMB by Land_type
## t = 0.24765, df = 14, p-value = 0.808
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.001592823  0.002008671
## sample estimates:
## mean in group A mean in group F
##     0.001883751     0.001675827
```

mean group F - mean group A =
-0.000207924

How do the equations look like?

Remember: $\overset{\text{regression function}}{\boxed{\mu(X)}} = \beta_0 + \beta_1 X$

$$\hat{\mu}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Estimated means of two level case:

$$\hat{\mu}_{\text{agriculture}}$$

$$\hat{\mu}_{\text{forest}}$$

Space of possible outcomes Ω = {agriculture, forest}.
Random variable X takes the values x = {0, 1}.

$$X(\omega) = \begin{Bmatrix} 0 & = & \text{agriculture} \\ 1 & = & \text{forest} \end{Bmatrix}$$

How do the equations look like?

$$X(\omega)=\begin{Bmatrix} 0 & = & \text{agriculture} \\ 1 & = & \text{forest} \end{Bmatrix}$$

Equations: $\hat{\mu}_{\text{agriculture}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0$

$$\hat{\mu}_{\text{forest}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1$$

Result from calculation:

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0018838  0.0005937   3.173  0.00677 **
## Land_typeF  -0.0002079  0.0008396  -0.248  0.80800
...
```

Equations: $\hat{\mu}_{\text{agriculture}} = 0.0018838 + 0$

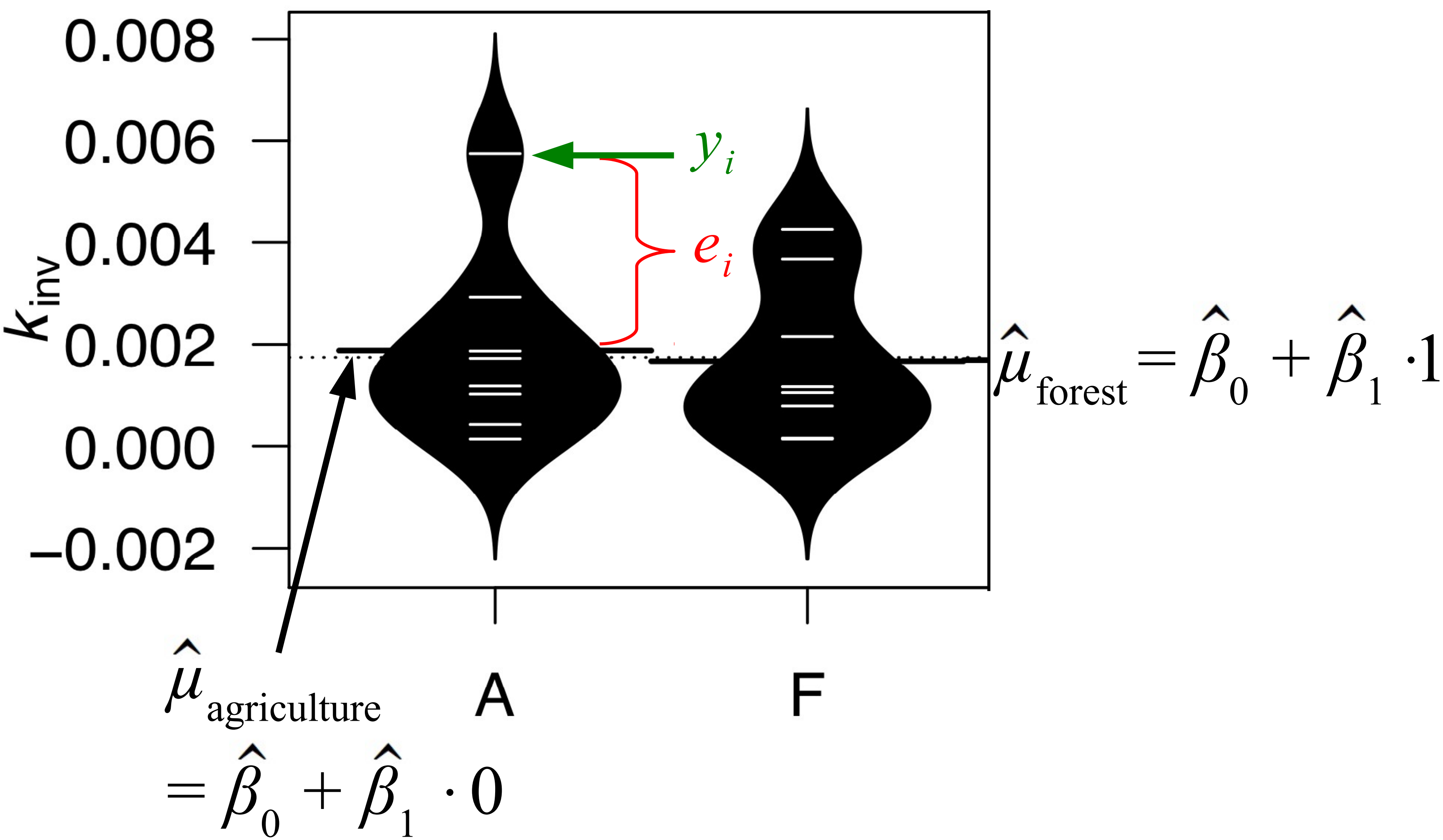
$$\hat{\mu}_{\text{forest}} = 0.0018838 - 0.0002079$$

Calculation of response y

Remember: $\boxed{y_i} = \boxed{b_0 + b_1 x_i} + \boxed{e_i}$ $\quad X(\omega) = \begin{Bmatrix} 0 & = & \text{agriculture} \\ 1 & = & \text{forest} \end{Bmatrix}$

Case i	ω	$\boxed{y}$ (from R)	$\boxed{\hat{\beta}_0 + \hat{\beta}_1 X}$	$\boxed{e}$ (from R)
1	Forest	0.00016	$= 0.00188 - 0.00021 +$	-0.00152
2	Forest	0.00427	$= 0.00188 - 0.00021 +$	0.00259
3	Forest	0.00079	$= 0.00188 - 0.00021 +$	-0.00088
4	Forest	0.00014	$= 0.00188 - 0.00021 +$	-0.00154
5	Forest	0.00367	$= 0.00188 - 0.00021 +$	0.00200
6	Forest	0.00216	$= 0.00188 - 0.00021 +$	0.00048
7	Forest	0.00117	$= 0.00188 - 0.00021 +$	-0.00051
8	Forest	0.00105	$= 0.00188 - 0.00021 +$	-0.00062
9	Agriculture	0.00043	$= 0.00188 + 0 +$	-0.00145
10	Agriculture	0.00102	$= 0.00188 + 0 +$	-0.00086
11	Agriculture	0.00173	$= 0.00188 + 0 +$	-0.00015
12	Agriculture	0.00187	$= 0.00188 + 0 +$	-0.00001
13	Agriculture	0.00575	$= 0.00188 + 0 +$	0.00387
14	Agriculture	0.00118	$= 0.00188 + 0 +$	-0.00070
15	Agriculture	0.00294	$= 0.00188 + 0 +$	0.00105
16	Agriculture	0.00014	$= 0.00188 + 0 +$	-0.00175

Graphical illustration of regression with a categorical predictor



Unity of the linear model and multiple linear regression

Contents

Categorical predictor in regression

Example with four-level predictor (land use example)

Mean Agriculture

p-values for pairwise comparisons with the Intercept (Agriculture)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0018838  0.0005660   3.328  0.00271 **
Land_typeF  -0.0002079  0.0008005  -0.260  0.79719
Land_typeU  -0.0001864  0.0008647  -0.216  0.83107
Land_typeV  -0.0001551  0.0008286  -0.187  0.85306
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001601 on 25 degrees of freedom
Multiple R-squared:  0.003219,  Adjusted R-squared:  -0.1164
F-statistic: 0.02691 on 3 and 25 DF,  p-value: 0.9939
```

Difference to intercept

Not matching (in contrast to two-level predictor)

F-statistic

Remember: MSS = Model Sum of Squares
RSS = Residual Sum of Squares
MSE = Mean Squared Error

In the context of the linear regression model with p predictors (or $p+1$ levels of a factor), the statistic is used to assess H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ (or H_0: $\mu_1 = \mu_2 = \ldots = \mu_{p+1}$).

The statistic is calculated as the division of two variances:

$$F_{(p,\, n-p-1)} = \frac{\left(\dfrac{\text{Explained var.}}{\boxed{\text{DoF model}}}\right)}{\left(\dfrac{\text{Unexplained var.}}{\boxed{\text{DoF error}}}\right)} = \frac{\left(\boxed{\dfrac{1}{p}}\text{MSS}\right)}{\left(\boxed{\dfrac{1}{n-p-1}}\text{RSS}\right)} = \frac{\left(\dfrac{\text{MSS}}{p}\right)}{\text{MSE}}$$

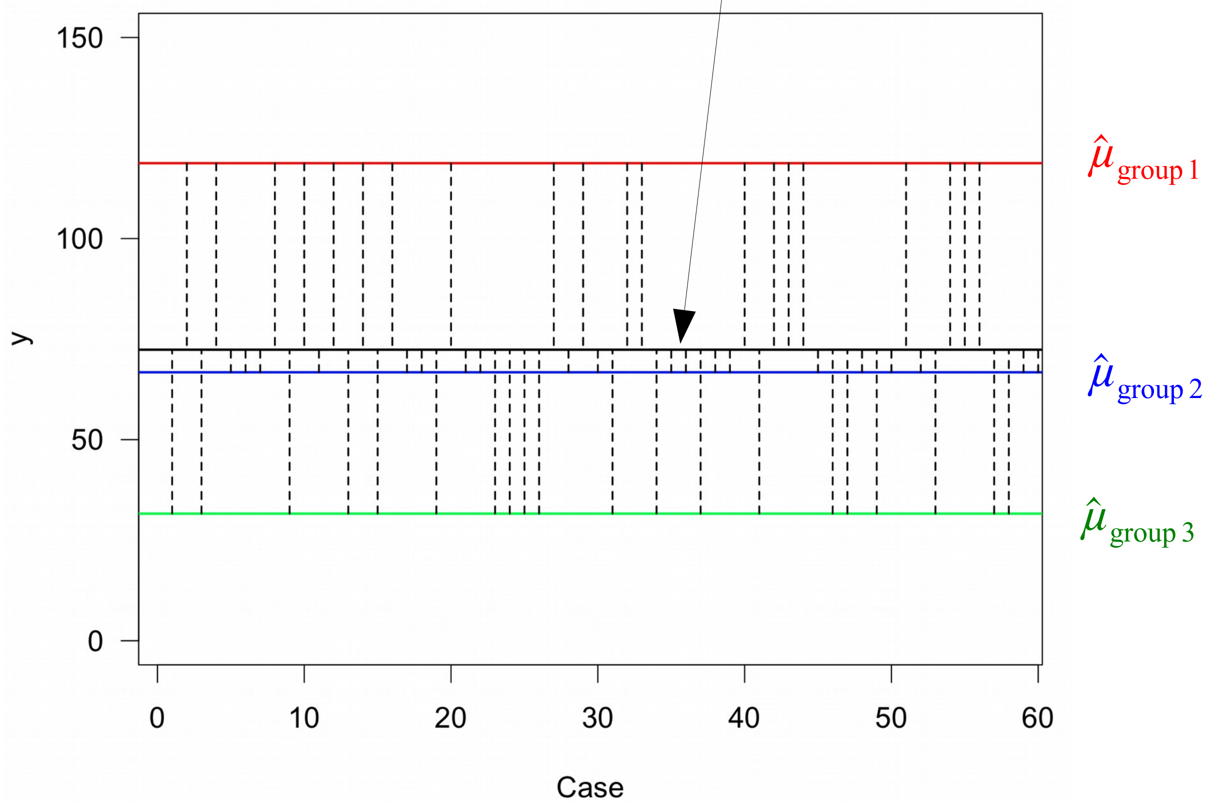
p = no. parameters (predictors) in model excluding intercept
n = sample size (no. of cases across all groups in case of factor)

Graphical illustration of F-statistic

$$F_{(p,\,n-p-1)} = \frac{\left(\dfrac{\text{Explained var.}}{\text{DoF model}}\right)}{\left(\dfrac{\text{Unexplained var.}}{\text{DoF error}}\right)} = \frac{\left(\dfrac{1}{p}\boxed{\text{MSS}}\right)}{\left(\dfrac{1}{n-p-1}\boxed{\text{RSS}}\right)} = \frac{\left(\dfrac{\text{MSS}}{p}\right)}{\text{MSE}}$$

$$\boxed{\text{MSS}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$\boxed{\text{RSS}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

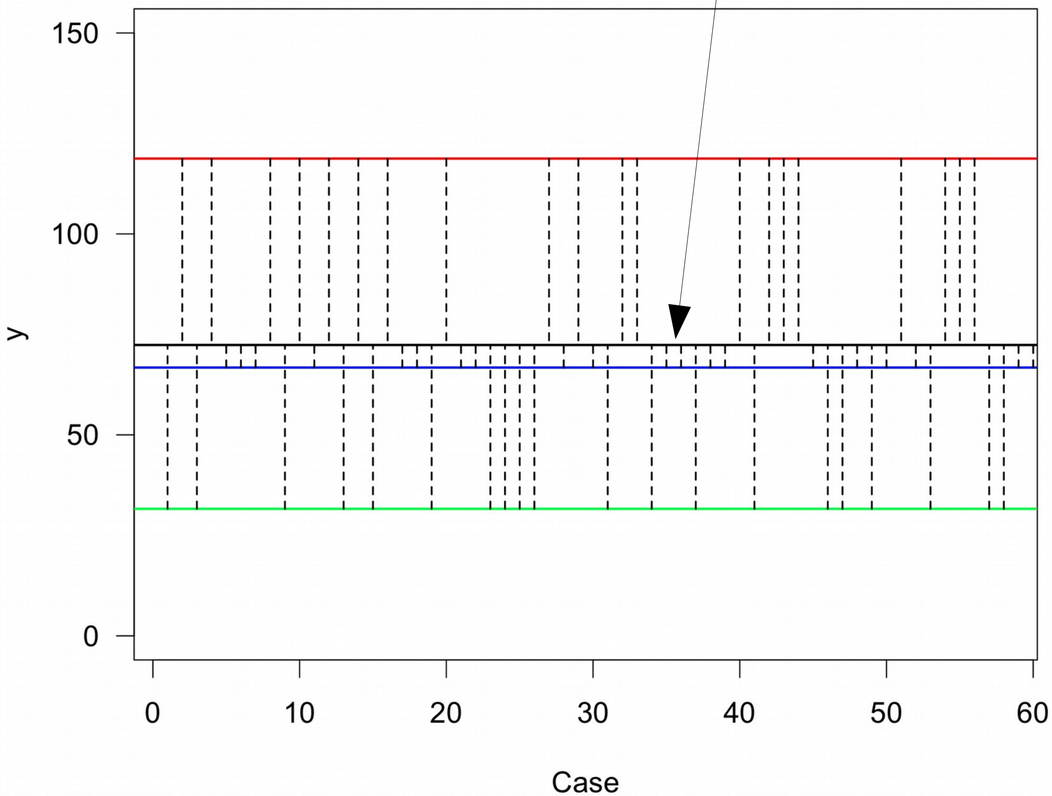


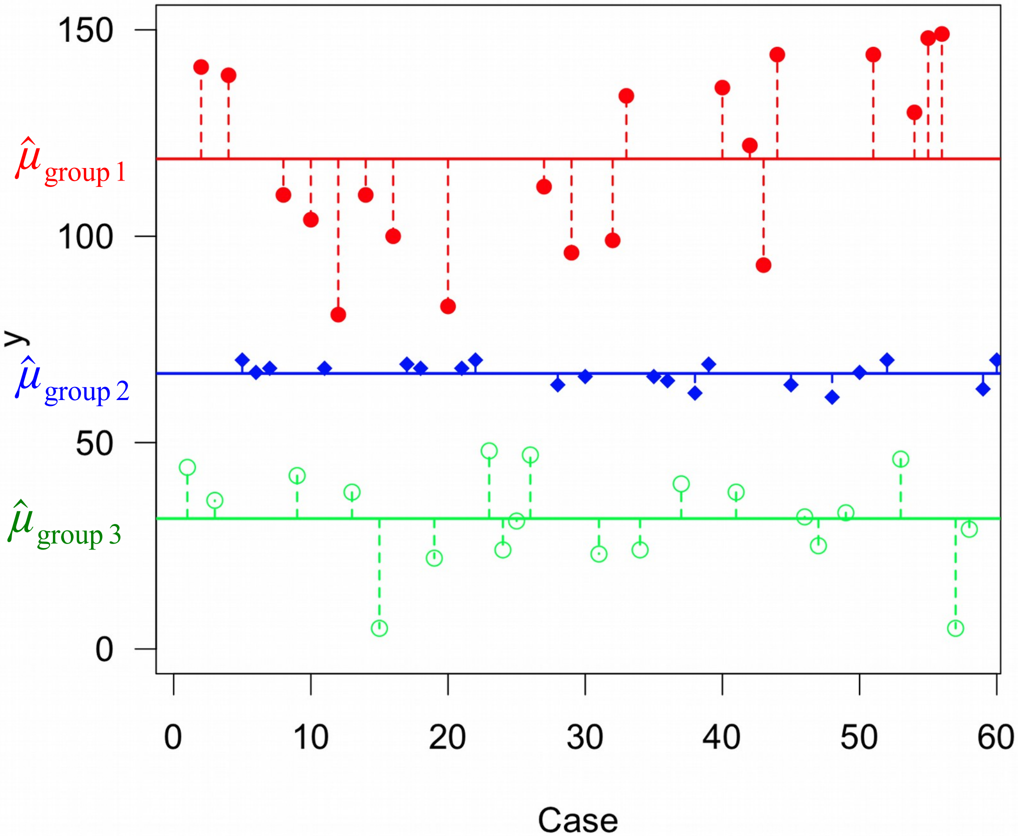
Graphical illustration of F-statistic

$$F_{(p,\,n-p-1)} = \frac{\left(\dfrac{\text{Explained var.}}{\text{DoF model}}\right)}{\left(\dfrac{\text{Unexplained var.}}{\text{DoF error}}\right)} = \frac{\left(\dfrac{1}{p}\boxed{\text{MSS}}\right)}{\left(\dfrac{1}{n-p-1}\boxed{\text{RSS}}\right)} = \frac{\left(\dfrac{\text{MSS}}{p}\right)}{\text{MSE}}$$

$$\boxed{\text{MSS}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$\boxed{\text{RSS}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$





Analysis Of Variance (ANOVA)

- ANOVA refers to test statistics (e.g. F-statistic) related to the question: Are j sample means drawn from (statistical) populations with equal μ?
- $H_0 : \mu_1 = \mu_2 = \ldots = \mu_j$ (The related alternative hypothesis would be: $H_{\mathrm{A}} : \exists i, j : \mu_i \neq \mu_j$)
- In case the ANOVA indicates different means, often multiple (pairwise) comparisons (e.g. with the t-test) are conducted afterwards to identify the means that differ. Multiple comparisons are a case of multiple inference.
- Multiple inference may require adjustment of p-values (if compared against a fixed threshold) and of confidence intervals.

Unity of the linear model

Regression

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0018838  0.0005660   3.328  0.00271 **
Land_typeF  -0.0002079  0.0008005  -0.260  0.79719
Land_typeU  -0.0001864  0.0008647  -0.216  0.83107
Land_typeV  -0.0001551  0.0008286  -0.187  0.85306
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.001601 on 25 degrees of freedom
Multiple R-squared:  0.003219,  Adjusted R-squared:  -0.1164
F-statistic: 0.02691 on 3 and 25 DF,  p-value: 0.9939
```

ANOVA with F-test

```
           Df    Sum Sq   Mean Sq F value Pr(>F)
Land_type   3 2.100e-07 6.900e-08   0.027  0.994
Residuals  25 6.408e-05 2.563e-06
```

Square-root

ANOVA assumptions and alternatives

- Given unity of the linear model:
  - Assumptions of regression model apply
  - Diagnosis differs slightly (e.g. residual vs. fitted plot suboptimal) → Use methods discussed for t-test
  - Dealing with violations → Approaches discussed for regression model largely applicable
  - As for regression model, violation of variance homogeneity assumption more problematic than of normal distribution assumption
- Alternative test statistics:
  - Welch's ANOVA for data with heterogeneous variances
  - Kruskal-Wallis test or permutational ANOVA for non-normally distributed data (if variance homogeneous!)


LET'S USE
MULTIPLE VARIABLES!
imgflip.com

Unity of the linear model and multiple linear regression

<u>Contents</u>

Linear model with multiple predictors

- Used to establish a (linear) relationship between multiple explanatory (predictor) and one response variable
- Model types:
  - Multiple continuous variables → Multiple linear regression
  - Multiple categorical variables → Multi-way ANOVA
  - Mix of continuous and categorical variables → ANCOVA
- Research goals same as for single-predictor linear model:
  - Prediction (predict response to multiple predictors)
  - Estimation (of regression coefficients and explained variance)
  - Assessing hypotheses (regarding relationship of response with multiple predictors)
  - Explanation (identify important explanatory variables)

Linear model with multiple predictors

Research goal: Explanation (identify important explanatory variables)

<u>Example:</u> Which variable(s) do best explain the response of different groups of organisms?

Table 2. Environmental Variables Selected in Linear Model Building with Highest Explanatory Power for the Response Variables Using Explained Variance (r^2) and the Akaike Information Criterion (AIC) as Goodness of Fit Measures

response variable	log mTU_{DM}	T (°C)	conductivity (μS/cm)	turbidity (NTU)	r^2	AIC
$SPEAR_{pesticides}$	x				0.67	−34
SIGNAL	x				0.36	98
bacteria[a]						
flagellates[a]		x	x		0.49	434
ciliates[a]		x		x	0.59	209
amoebas[a]				x	0.78	200

Schäfer et al. 2011

Multiple linear regression model

- Extension of simple linear regression model, we assume true relationship is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$

  $\rightarrow$ Classical definition for case i:

  $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \varepsilon_i$ with $\varepsilon \sim \text{Normal}(0, \sigma)$

- Using sample date, we estimate β's (b's = regression coefficients) to obtain estimates for y:

  $$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}$$

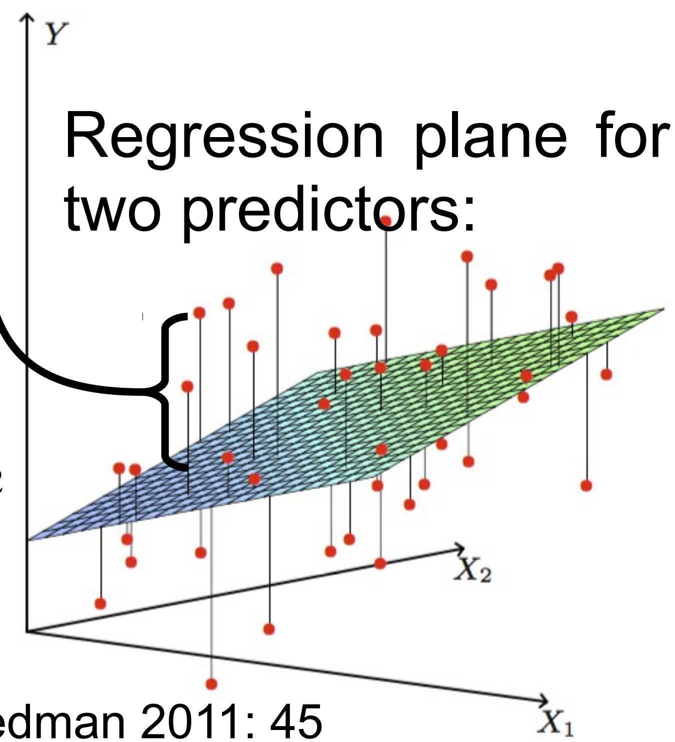
- Remember: Residual e_i defined as:

  $$e_i = y_i - \hat{y}_i$$

- Model fitting through minimising the squared sum of residuals (RSS):

$$\text{Find} \arg \min_{b_0, b_1, b_2, \ldots, b_p} \sum_{i=1}^{n} \left(y_i - (b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}) \right)^2$$

Regression plane for two predictors:



Taken from Hastie, Tibshirani and Friedman 2011: 45

Multiple linear regression model

Model in matrix form

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}$$

Notation for the number of observations (*n*):

$$\begin{aligned} \hat{y}_1 &= b_0 + b_1 x_{1,1} + b_2 x_{1,2} + \ldots + b_p x_{1,p} \\ \hat{y}_2 &= b_0 + b_1 x_{2,1} + b_2 x_{2,2} + \ldots + b_p x_{2,p} \\ &\vdots \\ \hat{y}_n &= b_0 + b_1 x_{n,1} + b_2 x_{n,2} + \ldots + b_p x_{n,p} \end{aligned} \quad \xrightarrow{\text{matrix}} \quad \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{X} b$$

Matrix calculations of regression coefficients same as explained for simple linear regression model

Interpretation of individual coefficients: Provide the effect size for one unit increase in the predictor on the response, if all other predictors remain constant.

Multi-way ANOVA

- Two-way ANOVA = Two categorical predictors, Three-way ANOVA = Three categorical predictors and so on
- Example: Effect of sex and population on possum length

Regression output

```
Call:
lm(formula = totlngth ~ sex * Pop, data = pos_dat)

Residuals:
     Min       1Q   Median       3Q      Max
-13.3333  -2.5227   0.5544   2.9773   9.4949

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  87.3684     0.9880  88.426   <2e-16 ***
sexm         -0.8633     1.2049  -0.716    0.475
PopVic        0.9649     1.3225   0.730    0.467
sexm:PopVic  -0.9473     1.7515  -0.541    0.590
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.307 on 100 degrees of freedom
Multiple R-squared:  0.03083,	Adjusted R-squared:  0.00175
F-statistic:  1.06 on 3 and 100 DF,  p-value: 0.3696
```

Mean of females in the other population

p-values for pairwise comparisons with Intercept (hypothesis: coefficient = 0)

Interaction term – discussed later

Square-root

ANOVA output

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	49.1	49.12	2.648	0.107
Pop	1	4.5	4.45	0.240	0.625
sex:Pop	1	5.4	5.43	0.293	0.590
Residuals	100	1854.8	18.55		

3

F-statistics differ!

Variance partitioning in multi-way ANOVA

lm(formula = totlngth ~ sex + Pop, data = pos_dat)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	49.1	49.12	2.648	0.107
Pop	1	4.5	4.45	0.240	0.625
sex:Pop	1	5.4	5.43	0.293	0.590
Residuals	100	1854.8	18.55		

lm(formula = totlngth ~ Pop + sex, data = pos_dat)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Pop	1	11.8	11.84	0.639	0.426
sex	1	41.7	41.72	2.250	0.137
Pop:sex	1	5.4	5.43	0.293	0.590
Residuals	100	1854.8	18.55		

Variance partitioning in multi-way ANOVA

lm(formula = totlngth ~ sex + Pop, data = pos_dat)

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS(sex)	sex	1	49.1	49.12	2.648	0.107
SS(Pop | sex)	Pop	1	4.5	4.45	0.240	0.625
SS(sex × Pop | sex, Pop)	sex:Pop	1	5.4	5.43	0.293	0.590
	Residuals	100	1854.8	18.55		

lm(formula = totlngth ~ Pop + sex, data = pos_dat)

		Df	Sum Sq	Mean Sq	F value	Pr(>F)
SS(Pop)	Pop	1	11.8	11.84	0.639	0.426
SS(sex | Pop)	sex	1	41.7	41.72	2.250	0.137
SS(sex × Pop | Pop, sex)	Pop:sex	1	5.4	5.43	0.293	0.590
	Residuals	100	1854.8	18.55		

- Difference due to Type 1 ANOVA, i.e. sequential SS, for unbalanced design
- Type 2: SS(sex | Pop) and SS(Pop | sex) as in Type 1 (→ sequence of model terms irrelevant), ignores interactions
- Type 3 considers interactions (e.g. SS (sex | Pop, sex × Pop))
- If aim is to assess hypotheses or partitioning of variance and interaction is irrelevant, use Type 2, otherwise Type 3.

Unity of the linear model and multiple linear regression

Contents

Interactions

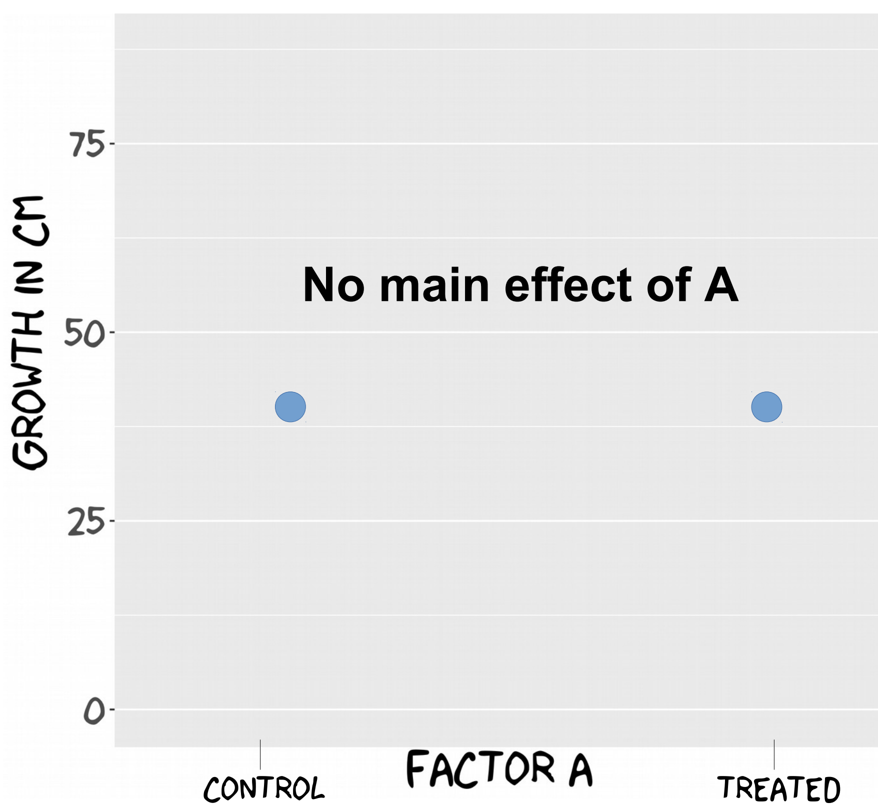
- Definition for linear model: Effect of one predictor on the response depends on state of another predictor

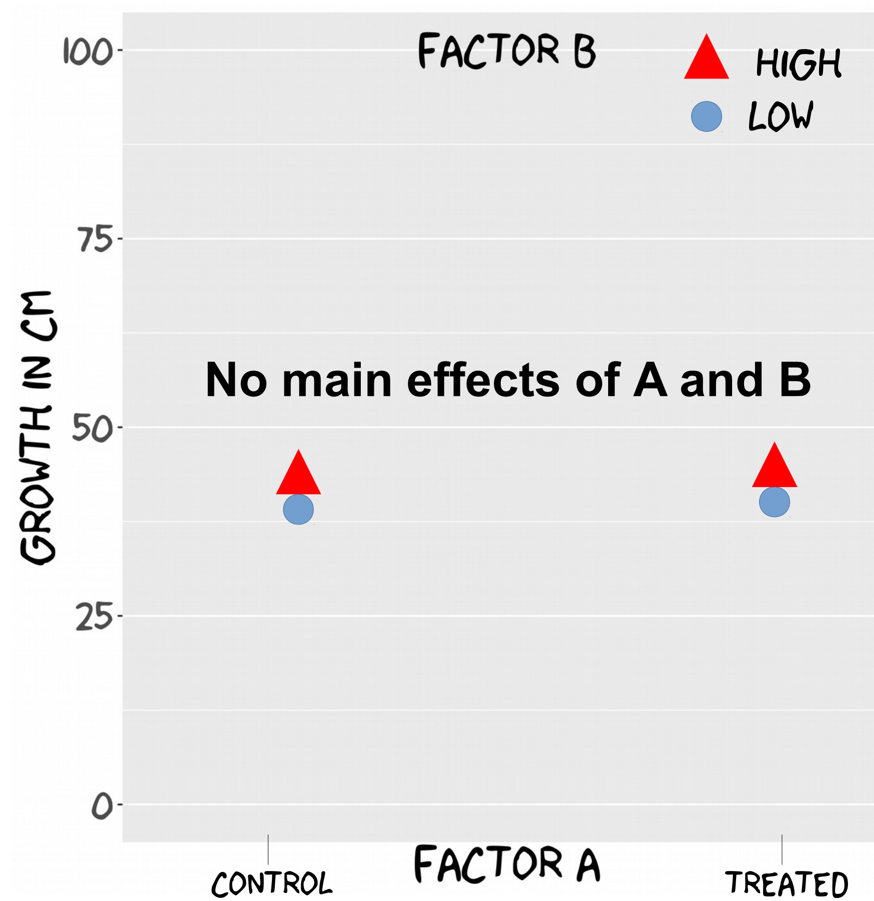
- Notation for two variables: $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,2}$
  Common short notation: $\mathrm{A} + \mathrm{B} + \mathrm{A} \times \mathrm{B}$

Main effects

Interaction effect

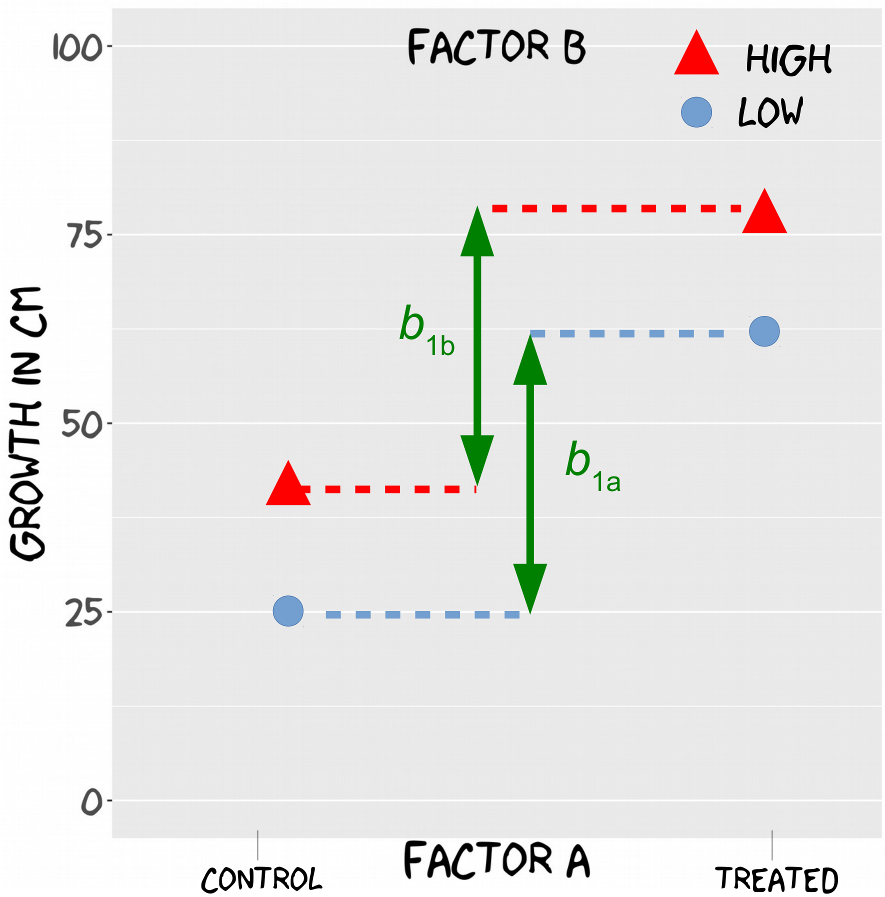
Illustration:

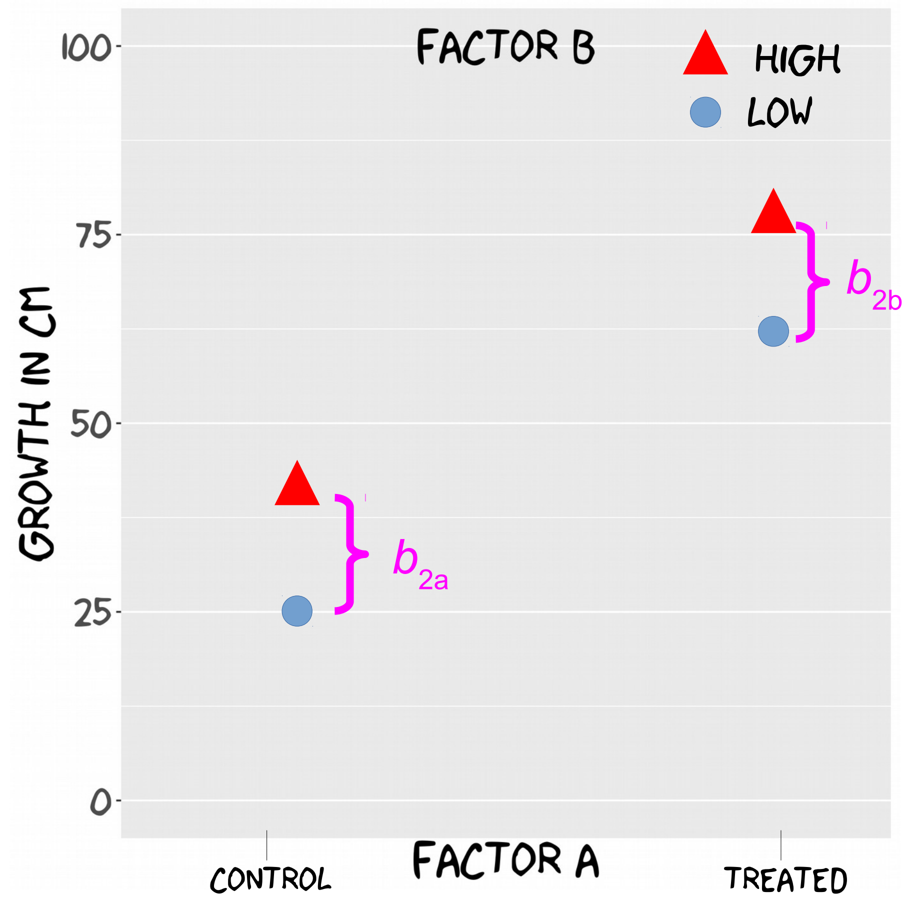




Interactions: Illustration

Main effects of A and B, no interaction





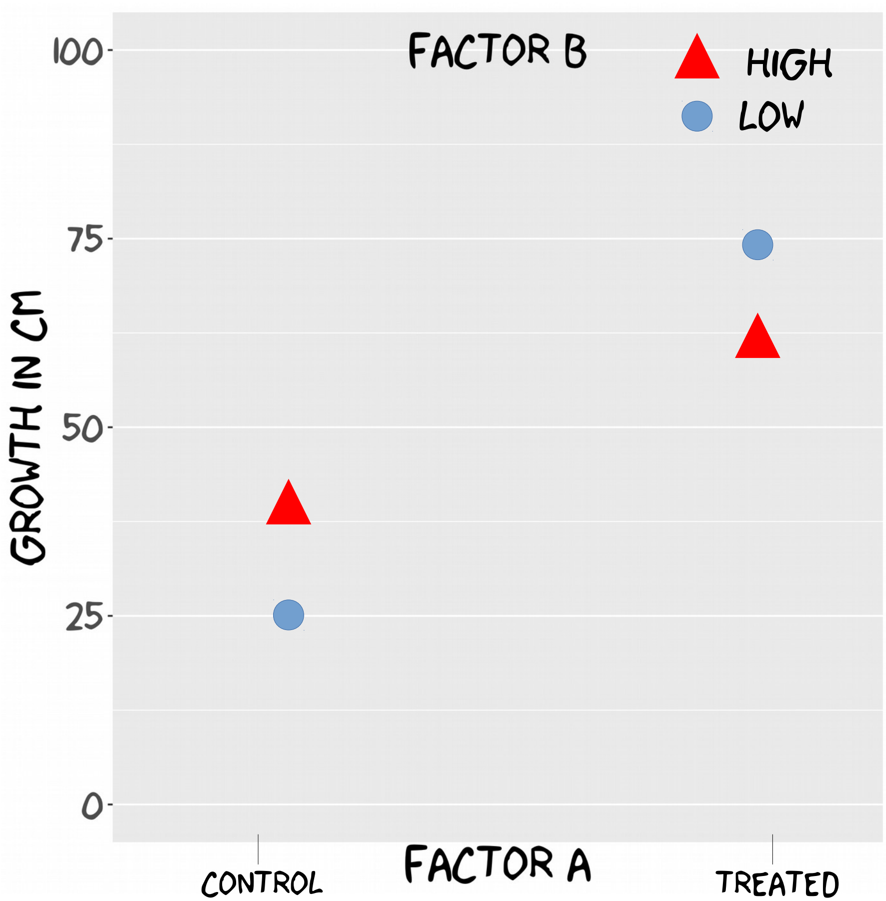
Effect of A: $$b_1 = \frac{b_{1a} + b_{1b}}{2}$$

Effect of B: $$b_2 = \frac{b_{2a} + b_{2b}}{2}$$

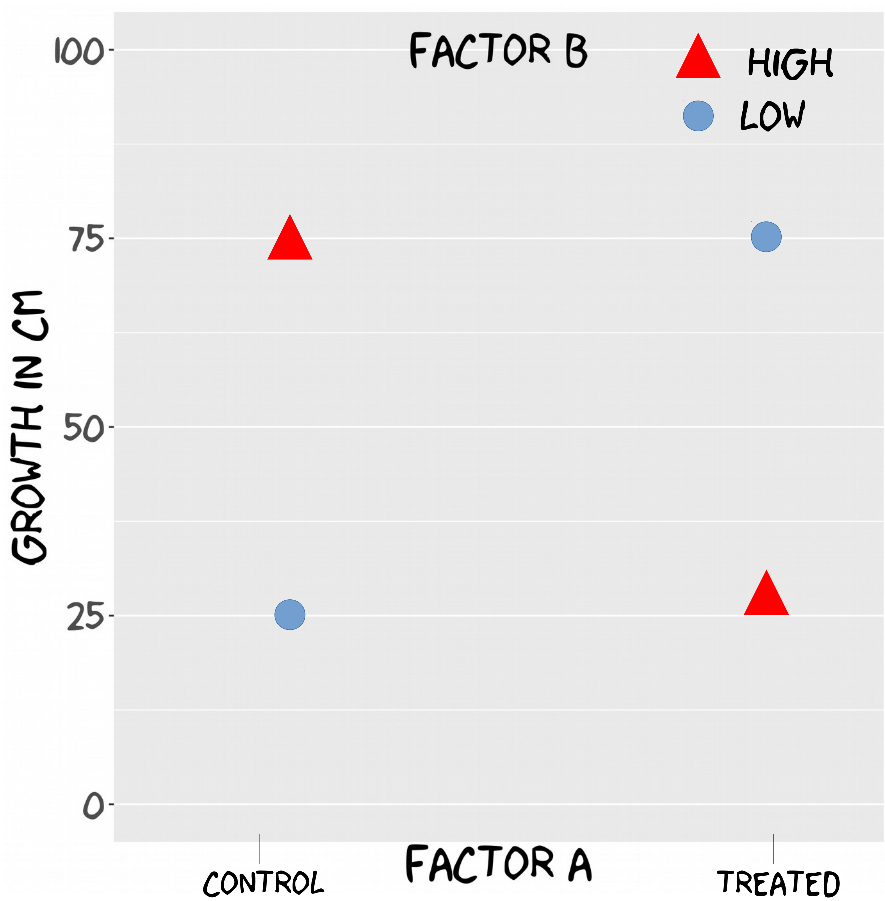
Total effect is additive (Control & Low compared to Treated & High) = $b_{1a} + b_{2b}$

Interactions: Illustration

Main effect of A, minor interaction effect, no B effect

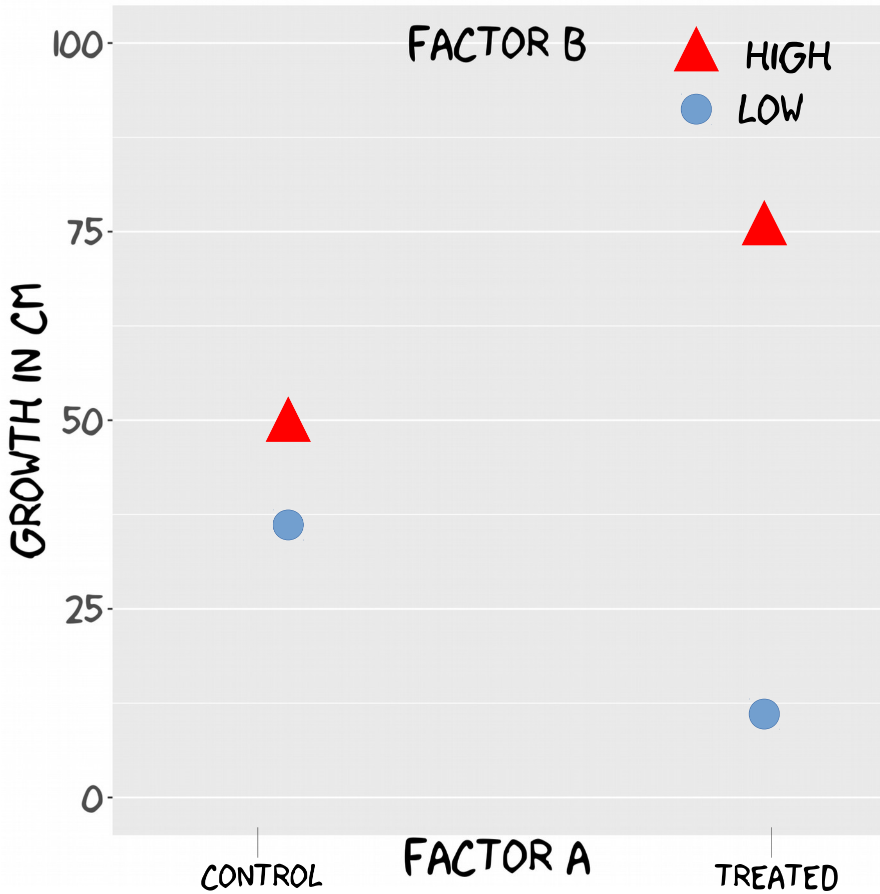


No main effects, interaction effect

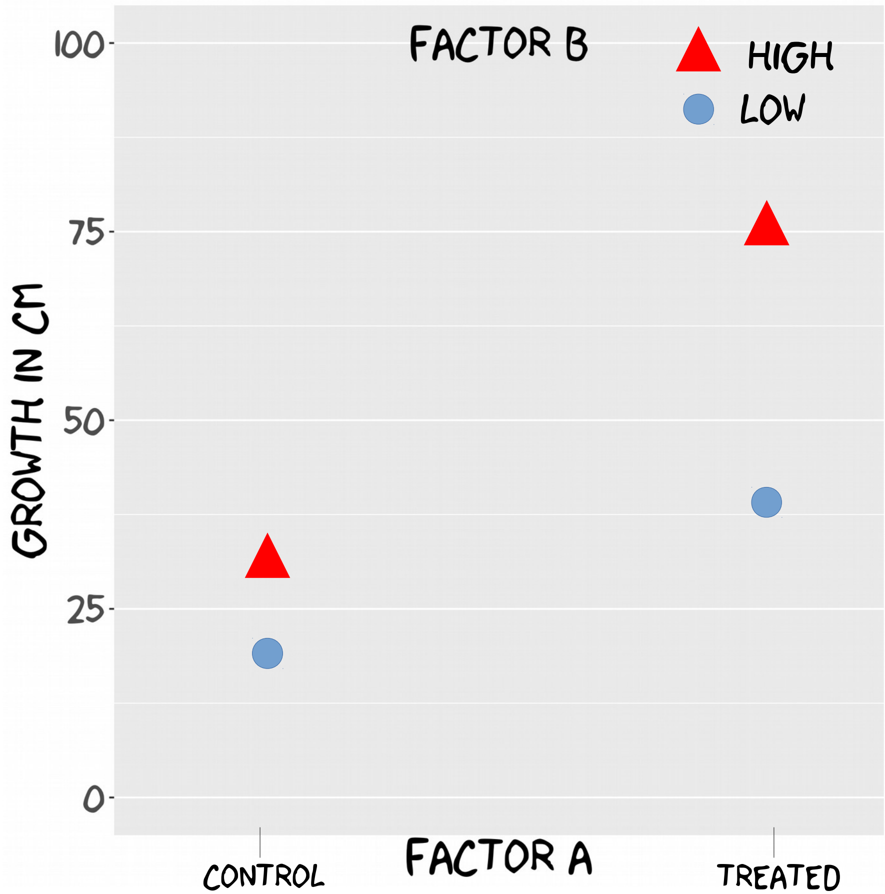


Interactions: Illustration

Main effect of B, large interaction effect, no A effect



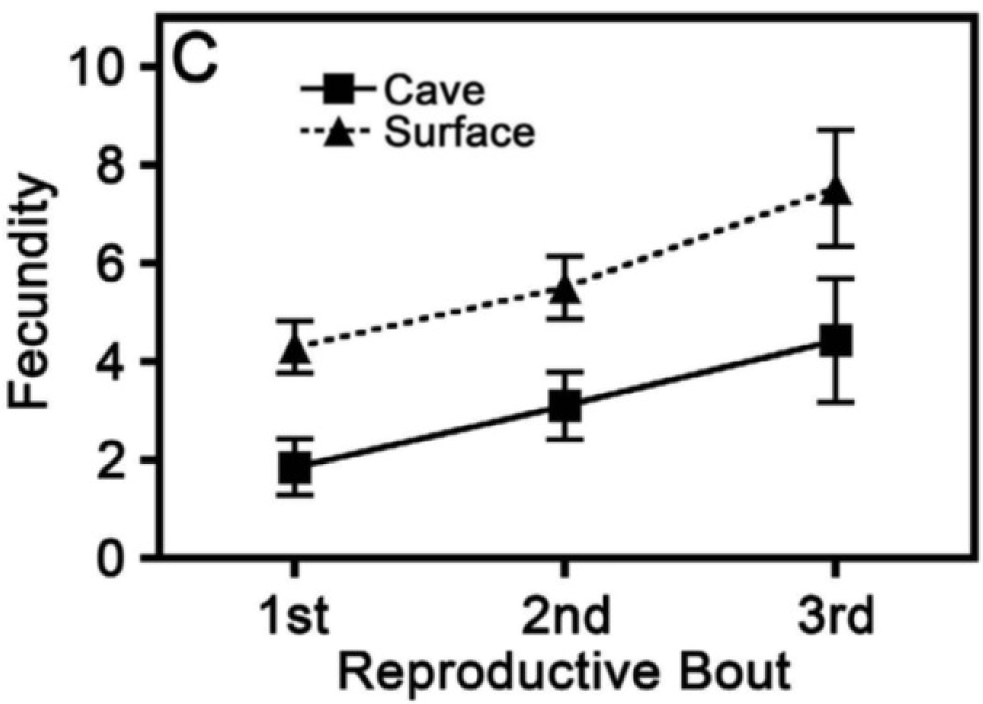
Main effect of A and B, interaction effect



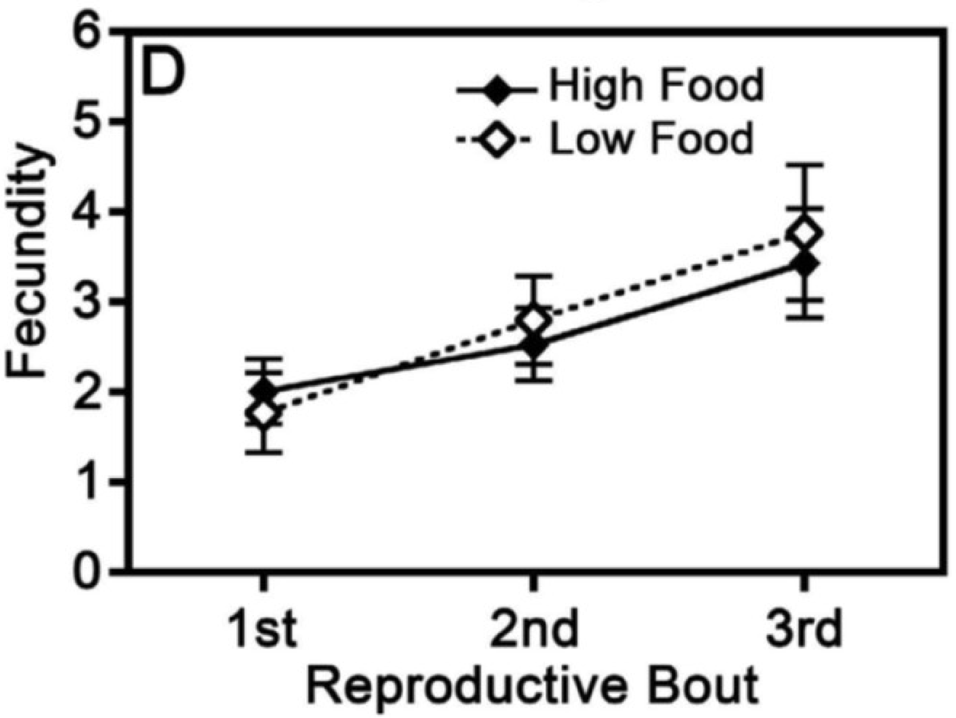
Interactions: Real world example

Study on the adaptation of mollis from different origins (caves or surface water)

Main effect of origin and reproductive bout, no interaction

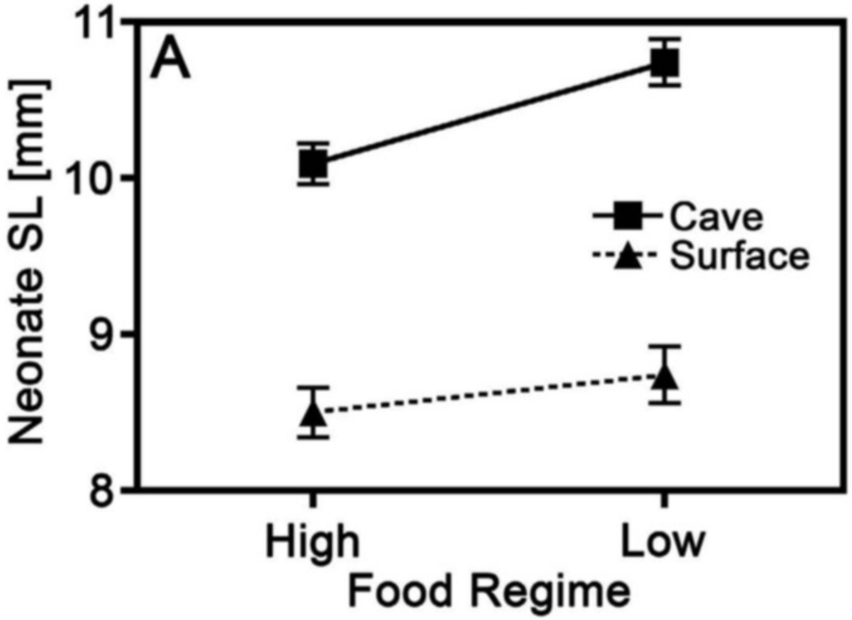


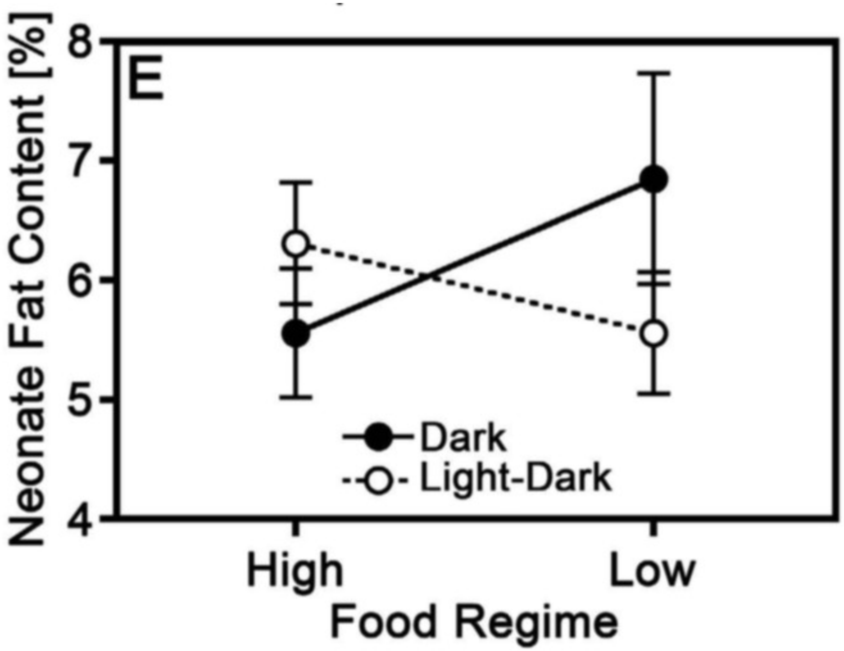
Main effect of reproduction, no interaction effect and of food

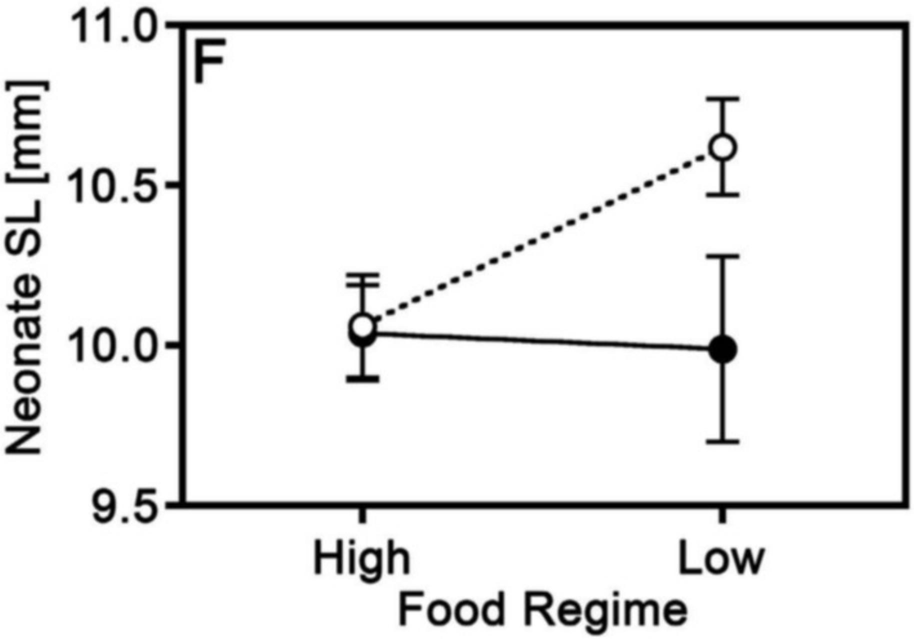


Riesch et al. (2016) *Sci. Rep. 6*

Interpret the effects including interactions!







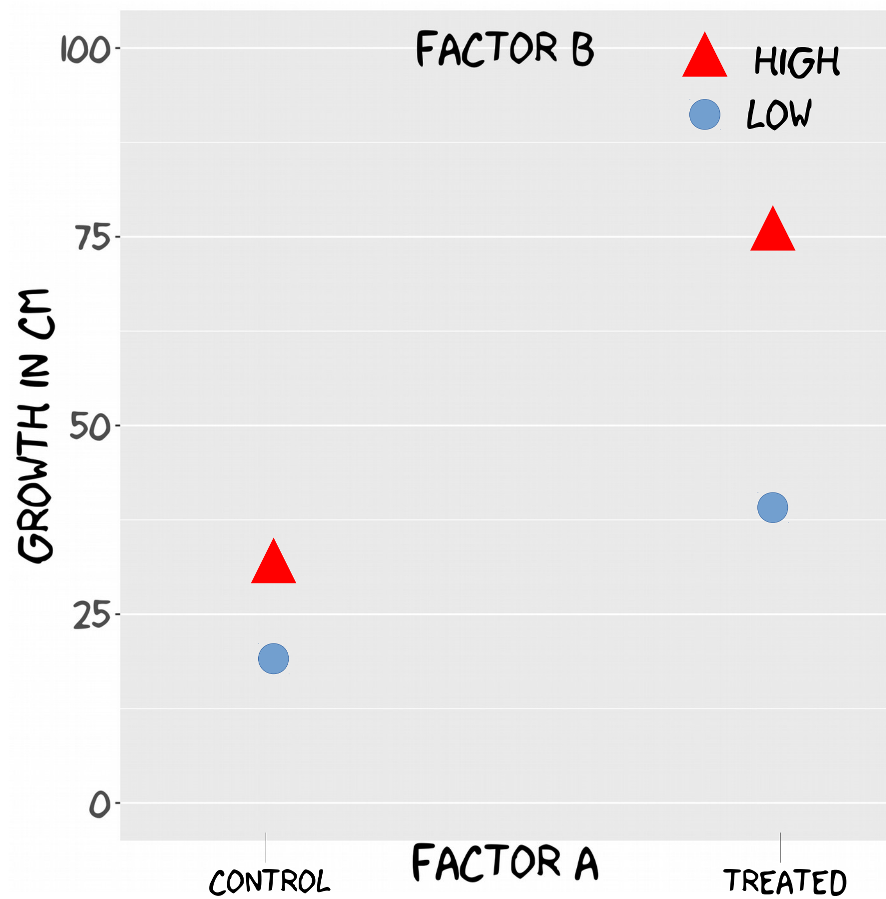
Riesch et al. (2016) *Sci. Rep. 6*

Interpreting main effects in presence of interactions

When is an interaction present?

- Classic approach: p-value < 0.05
- Visual inspection

Illustration:



Simplified, idealized R output

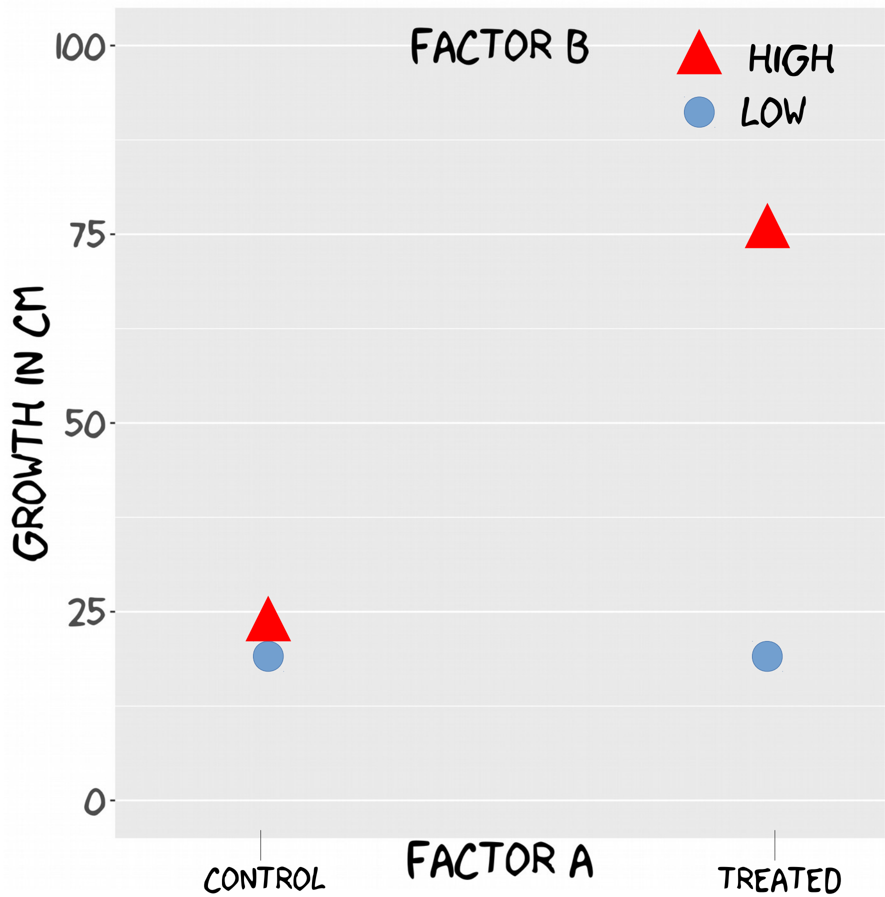
ANOVA (summary.aov)

	Pr(>F)
Factor 1	<0.001
Factor 2	<0.001
Factor1:Factor2	<0.001

Main effect of A and B can be interpreted as increasing growth, despite interaction effect

Interpreting main effects in presence of interactions

Illustration:



Simplified, idealized R output

ANOVA (summary.aov)

	Pr(>F)
Factor 1	<0.001
Factor 2	<0.001
Factor1:Factor2	<0.001

Interpretation of main effects of A and B is misleading

Final notes on interpretation of main effects in presence of interactions

- If Sum of Squares and coefficients of interaction term(s) << main terms, interaction likely irrelevant
- Interpret only after visual checking
- Multiple linear regression: Standardize data before including interaction terms:
  - Reduces collinearity with main variables
  - Improves interpretation

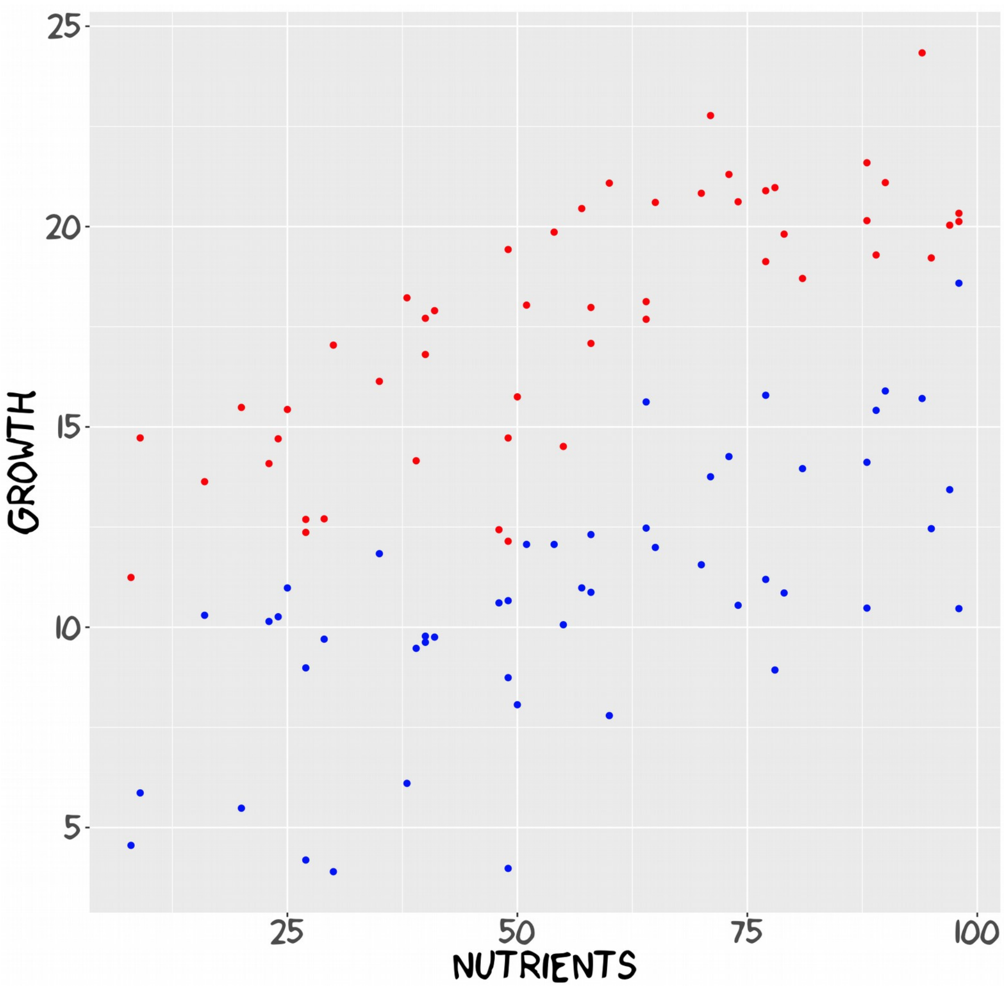
Unity of the linear model and multiple linear regression

Contents

Analysis Of Covariance (ANCOVA)

- Combination of continuous and categorical predictors

<u>Experiment: Response of plant growth to nutrients in two soils</u>



We could fit data to two independent regression models:

$$Y_{\text{Clay}} = \beta_{0,\ \text{Clay}} + \beta_{1,\ \text{Clay}} X_1 + \varepsilon$$

$$Y_{\text{Loam}} = \beta_{0,\ \text{Loam}} + \beta_{1,\ \text{Loam}} X_1 + \varepsilon$$

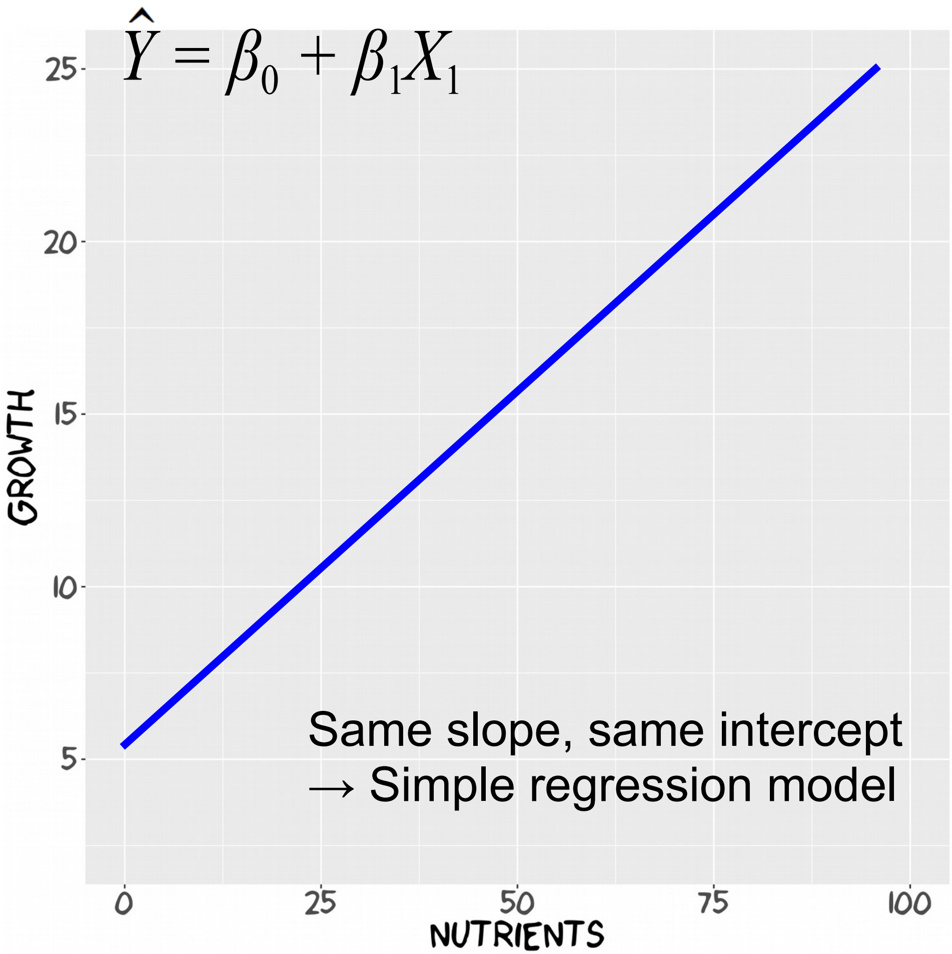
But what about a single model?

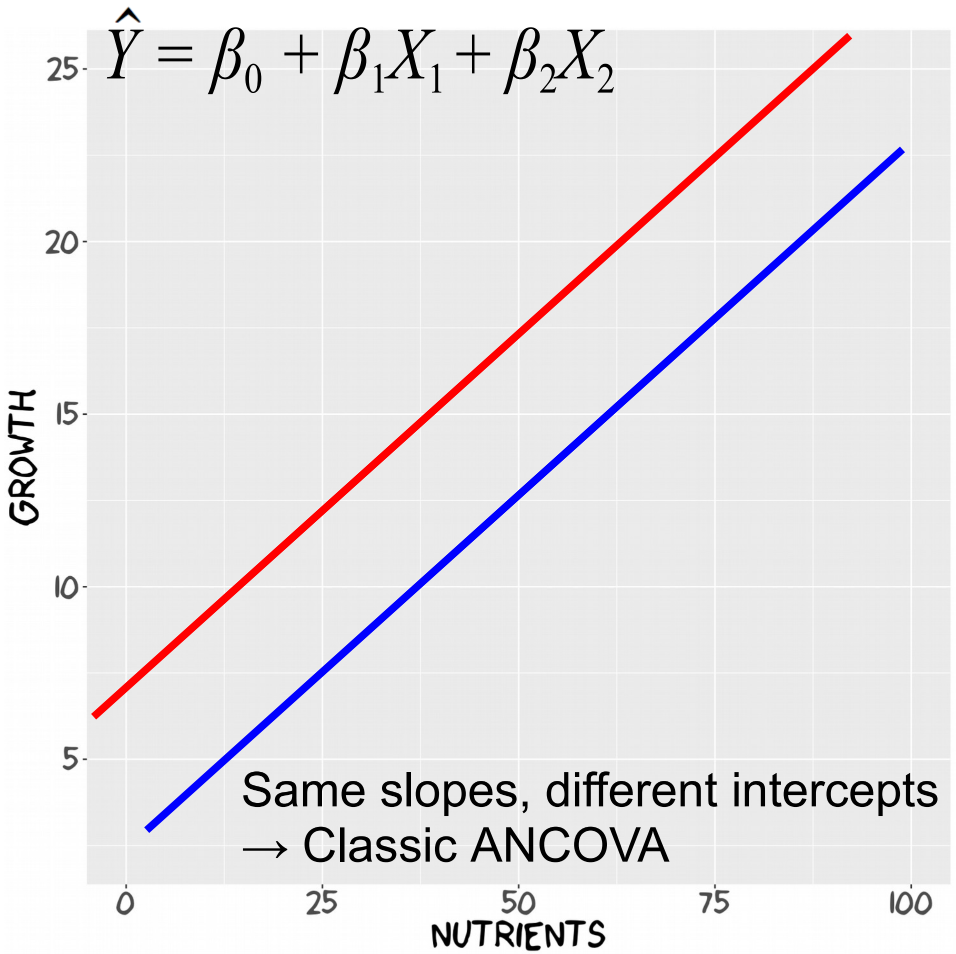
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$X_2(\omega) = \begin{Bmatrix} 0 & = & \text{Loam} \\ 1 & = & \text{Clay} \end{Bmatrix}$$

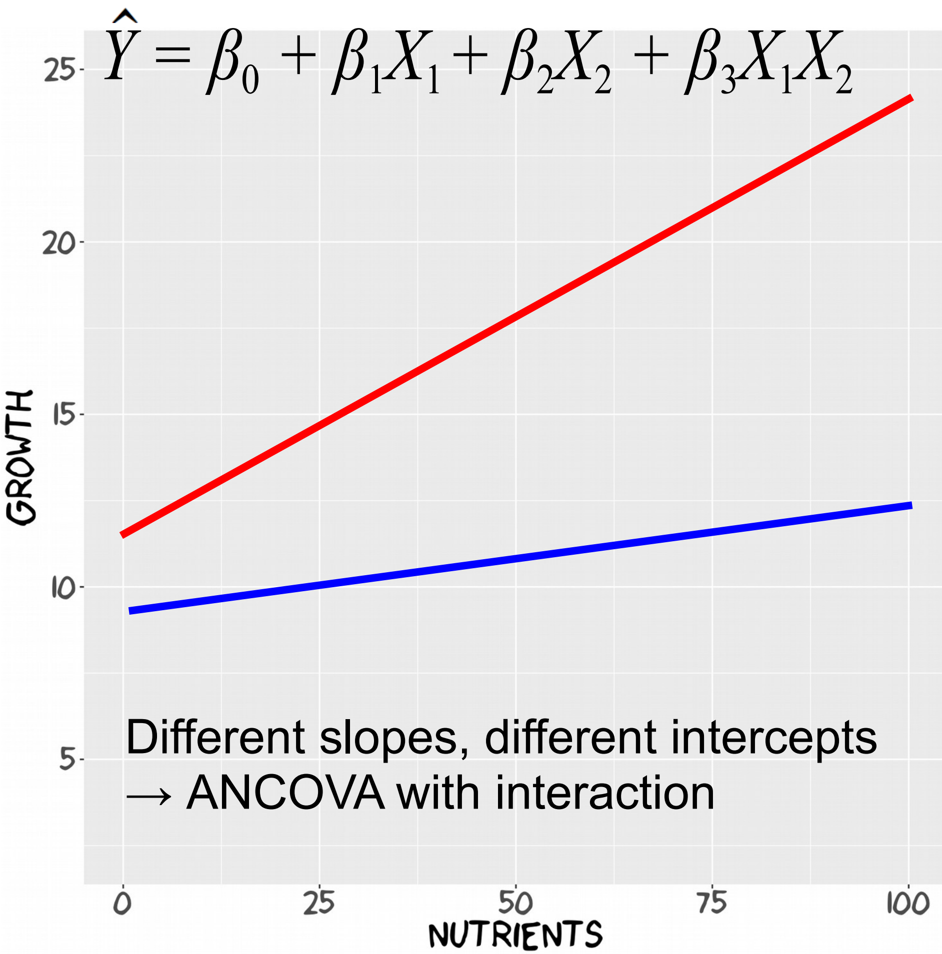
Analysis Of Covariance (ANCOVA)

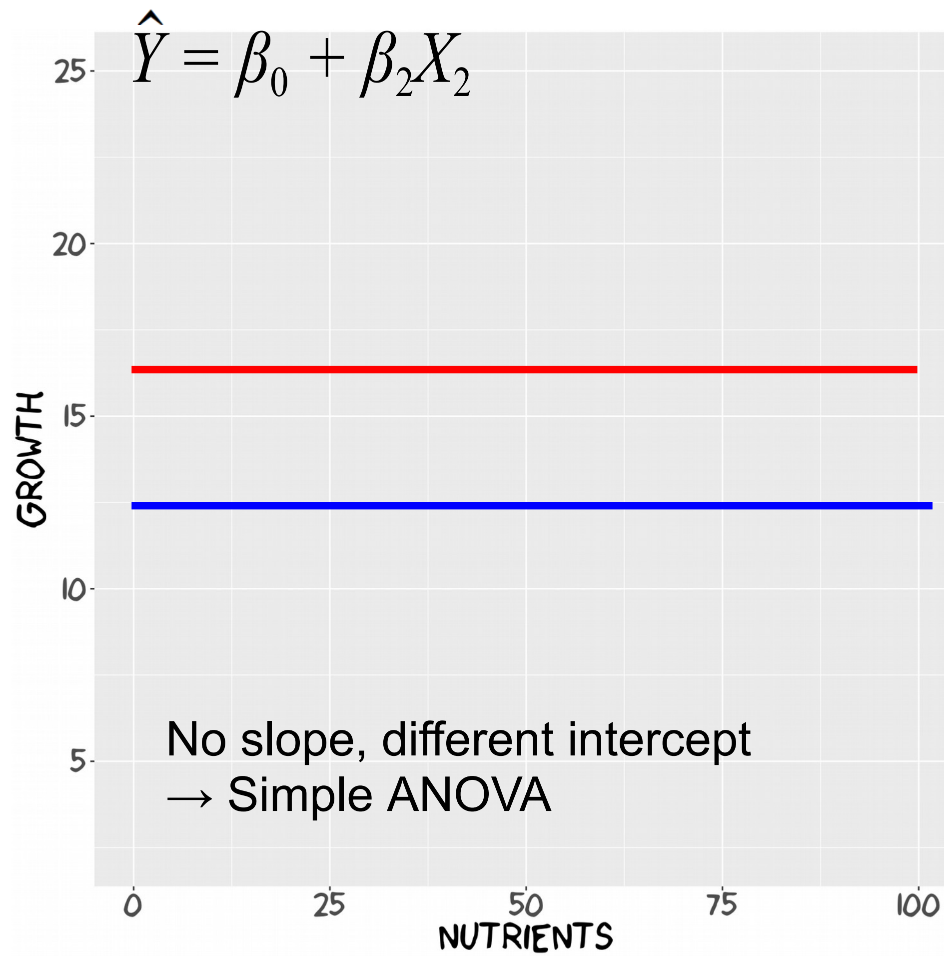
- Advantage of single model:
  - Less parameters compared to independent models
  - More comprehensive interpretation
  - Allows for assessment of different hypotheses (e.g. slopes are equal, intercept equal)



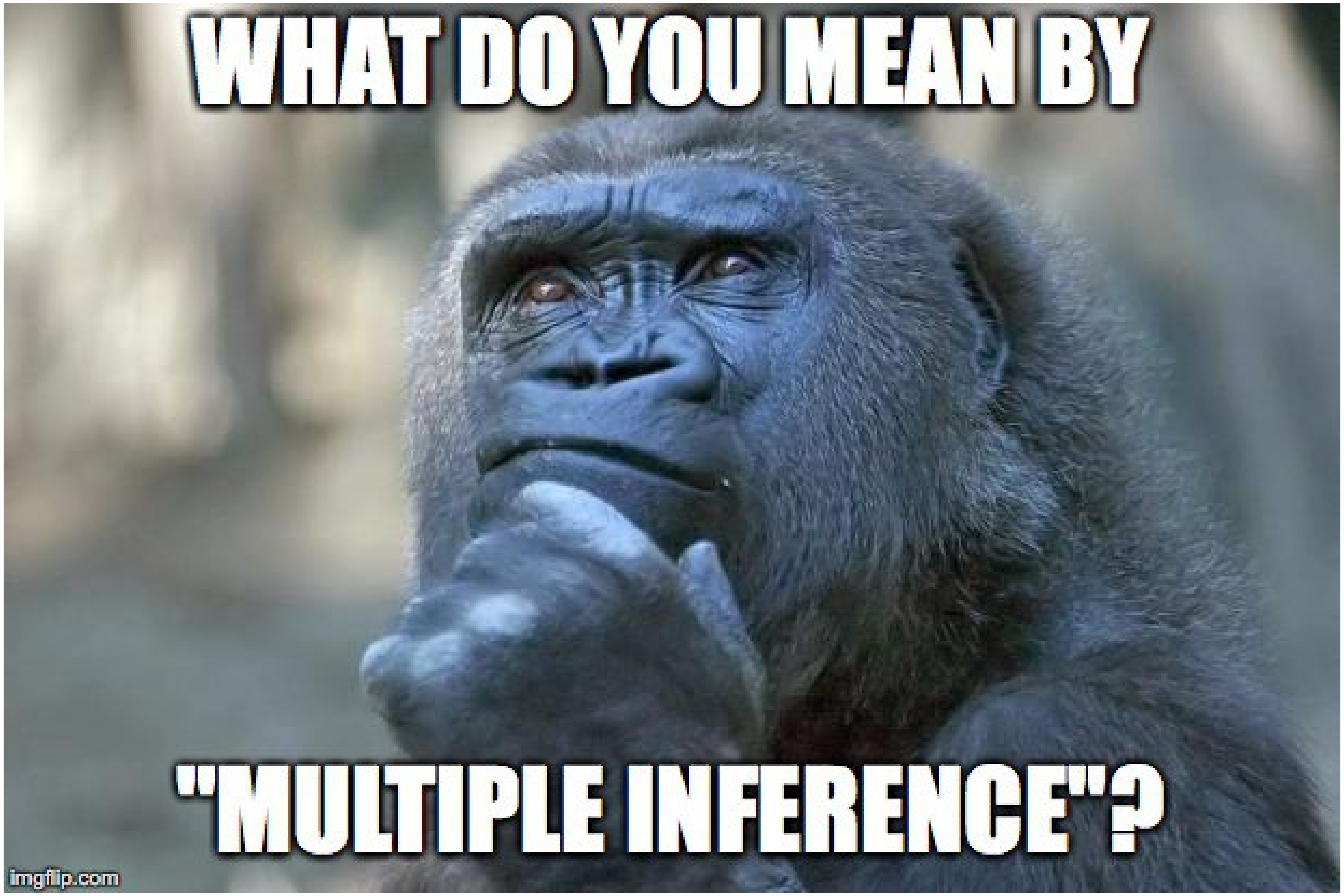


Analysis Of Covariance (ANCOVA)





- How to decide? Sequential assessment of hypotheses: 1. $\beta_3 = 0$? Yes → $2. \beta_2 = 0$ or $\beta_1 = 0$?
- Issue of multiple inference


WHAT DO YOU MEAN BY
"MULTIPLE INFERENCE"?
imgflip.com

Unity of the linear model and multiple linear regression

Contents

Multiple inference: CIs

Background: If computing 100 CIs, 5 will not include the true parameter. If we will calculate several CIs from future study data what is the probability that one of them excludes the true value?

1 CI: $1-(0.95) = 0.05$

2 CIs: $1-(0.95)^2 = 0.1$

…

5 CIs: $1-(0.95)^5 = 0.23$

If the objective is to keep the probability across all CIs at 0.05 (called family-wise error rate), you need to adjust the confidence level.

Example for adjustment: 5 CIs: $1-(0.99)^5 = 0.05$

Multiple inference: p-values

- Same issue when assessing hypotheses using a fixed significance threshold (i.e. following PaleoFisherian approach using a significance level α = 0.05)
- When do I have to adjust?
  - Difference between selective and simultaneous inference.
    - Selective inference: Inference on a subset of data that has been manually or automatically selected in the light of data
    - Simultaneous inference: Inference on all possible relationships or comparisons in the data set
  - In case of selective inference and in case of simultaneous inference, if parts of data re-used in statistical procedure
  - Not in case of simultaneous inference, if statistical procedures (e.g. specific hypotheses for assessment) and data pre-processing have been defined *a priori* and data (re-)used is independent (e.g. orthogonal contrasts)
  - Not consistently handled in scientific literature

Multiple inference: How to adjust?

- How to adjust if I have to adjust?
  - Many approaches for adjusting p-values (for details and implementation in R see Bretz et al. 2011)
  - Classical approach: Correct for number of statistical procedures (e.g. tests) using, for example, Bonferroni correction (focus is on family-wise error rate)
  - More recent approach: Control the false discovery rate (rate of falsely rejected null hypotheses). Involves the stepwise adjustment of p-values (or CIs)
  - Alternative: Move from hypothesis testing (particularly in the Neyman-Pearson or PaleoFisherian paradigm) to parameter estimation (e.g. with Bayesian models (Gelman et al. 2012))

Tools for complex data analysis

University of Koblenz-Landau 2018/19

Unity of the linear model and multiple linear regression

Ralf B. Schäfer

These slides and notes complement the lecture with exercises "Tools for complex data analysis" for ecotoxicologists and environmental scientists. Do not hesitate to contact me if you have any comments or you find any errors (slides, slide notes, or code): schaefer-ralf@uni-landau.de

While I made notes below the slides, some aspects are only mentioned in the R demonstration associated with the lecture.

Learning targets

- Describe the mathematical basis for including categorical predictors
- Explain and interpret analysis of variance
- Explain and interpret linear models with multiple predictors
- Describe the issue of multiple inference

Learning targets

- Describe the mathematical basis for including categorical predictors
  - Outline the relation between the t-test or an ANOVA with one variable and a linear regression with a categorical predictor.
  - Interpret the regression coefficients for categorical predictors.
- Explain and interpret analysis of variance
  - For which research questions can you use ANOVA?
  - Describe the concept of sum of squares for ANOVA. How are they calculated?
  - What are the assumptions of ANOVA and how can you check them?

Learning targets

- Explain and interpret linear models with multiple predictors
  - Classify the types of linear models with multiple predictors.
  - Why does the F-statistic differ between linear regression and ANOVA output?
  - Explain the types of ANOVA and when they should be used.
  - What are main effects and interactions?
  - How do interactions complicate the interpretation of main effects?
  - Why should you standardize variables before multiple linear regression if interactions are present?
- Describe the issue of multiple inference
  - What is the issue of multiple inference?
  - When and how do you need to account for multiple inference?

Unity of the linear model and multiple linear regression

Contents

- **Categorical predictors in regression analysis**
- Analysis of variance with *F*-test and unity of the linear model
- Linear model with multiple predictors
- Interactions
- Analysis of covariance
- Multiple inference

Case study: Land use

<u>Research question:</u> Does land use influence the invertebrate leaf breakdown rate k_{inv} in streams?

<u>Study:</u> 29 streams with four major land uses in upstream catchment examined for leaf breakdown.

6 Pictures: Colourbox.de or own

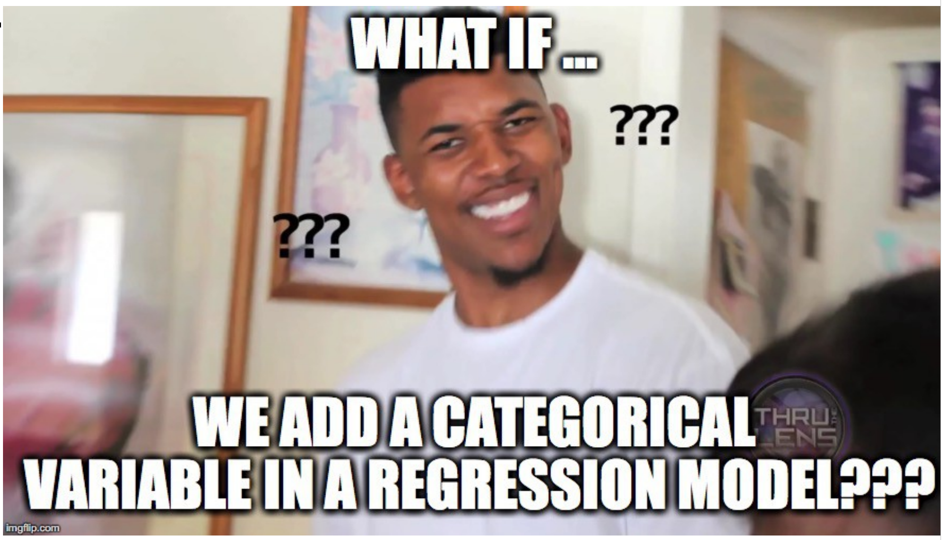
Leaf breakdown is an important ecosystem process providing energy to the stream ecosystem. The invertebrate leaf breakdown rate was examined in 29 streams across different land uses in Southern Rhineland Palatinate. More details on the study can be found in:

Voß K. & Schäfer R.B. (2017) Taxonomic and functional diversity of stream invertebrates along an environmental stress gradient. Ecological Indicators 81, 235–242.

Case study: Land use

<u>Research question:</u> Does land use influence the invertebrate leaf breakdown rate k_{inv} in streams?

<u>Required analysis:</u> Relate k_{inv} (cont.) to land use type (categ.).

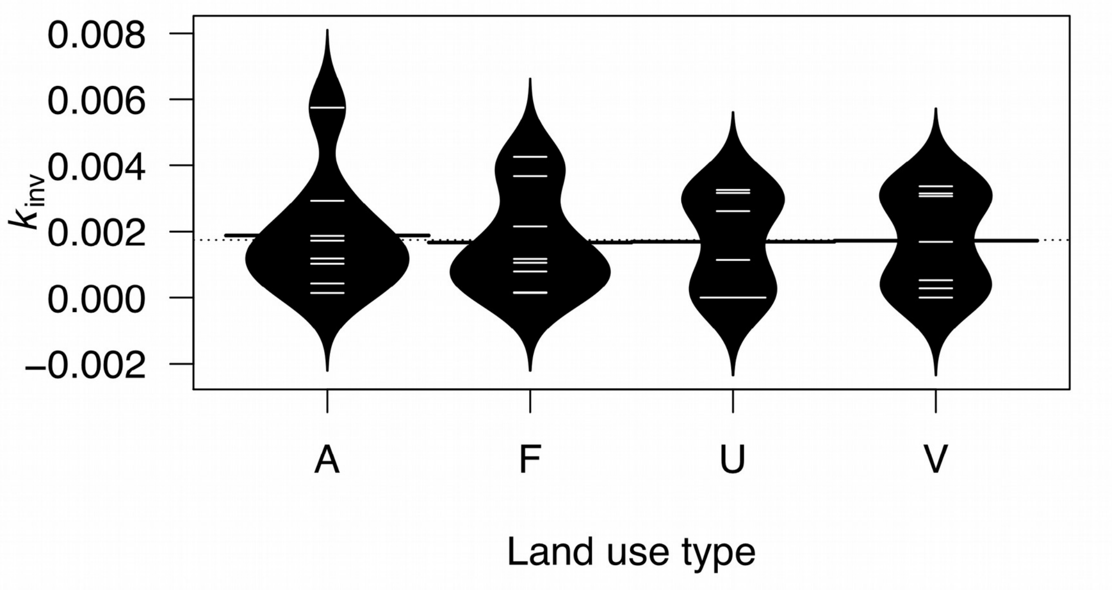


7

To answer the research question, we need to relate the variable land use type, which is a categorical variable, to the invertebrate leaf breakdown rate k_m.

Case study: Land use

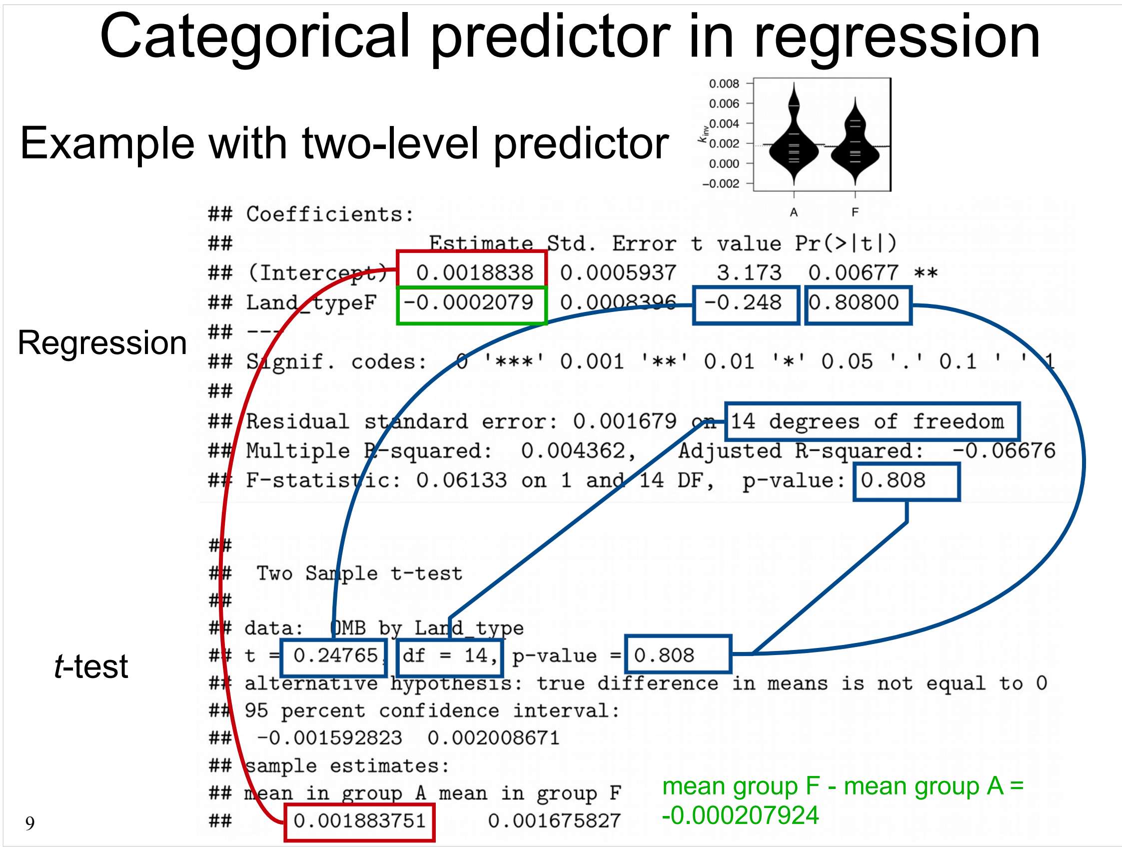
Graphical relationship of k_{inv} with land use type.



8

A = agriculture, F = forest, U = urban, V = viniculture

All groups have a similar variance and the mean looks similar across groups.



A = agriculture, F = forest

For didactical reasons, we initially restrict our data set to two levels.

All groups have a similar variance and the mean looks similar across groups.

Note that the *p*-values related to the *F*-statistic and *t*-statistic are the same in the case of categorical predictors with two levels. This will not be the case with more than two levels as we will see later.

For further details see the R tutorial related to this session.

How do the equations look like?

regression function

Remember: $\boxed{\mu(X)} = \beta_0 + \beta_1 X$

$$\hat{\mu}(X) = \hat{\beta}_0 + \hat{\beta}_1 X$$

Estimated means of two level case:

$$\hat{\mu}_{\text{agriculture}}$$

$$\hat{\mu}_{\text{forest}}$$

Space of possible outcomes Ω = {agriculture, forest}.
Random variable X takes the values x = {0, 1}.

$$X(\omega) = \begin{Bmatrix} 0 & = & \text{agriculture} \\ 1 & = & \text{forest} \end{Bmatrix}$$

How do the equations look like?

$$X(\omega)=\begin{Bmatrix} 0 & = & \text{agriculture} \\ 1 & = & \text{forest} \end{Bmatrix}$$

Equations: $\hat{\mu}_{\text{agriculture}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0$

$$\hat{\mu}_{\text{forest}} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1$$

Result from calculation:

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0018838  0.0005937   3.173  0.00677 **
## Land_typeF  -0.0002079  0.0008396  -0.248  0.80800
```

Equations: $\hat{\mu}_{\text{agriculture}} = 0.0018838 + 0$

$$\hat{\mu}_{\text{forest}} = 0.0018838 - 0.0002079$$

Compared to a regression model, we only have two different fitted response values, i.e. the estimated means.

Calculation of response y

Remember: $\boxed{y_i} = \boxed{b_0 + b_1 x_i} + \boxed{e_i}$ $\qquad X(\omega) = \begin{cases} 0 = \text{agriculture} \\ 1 = \text{forest} \end{cases}$

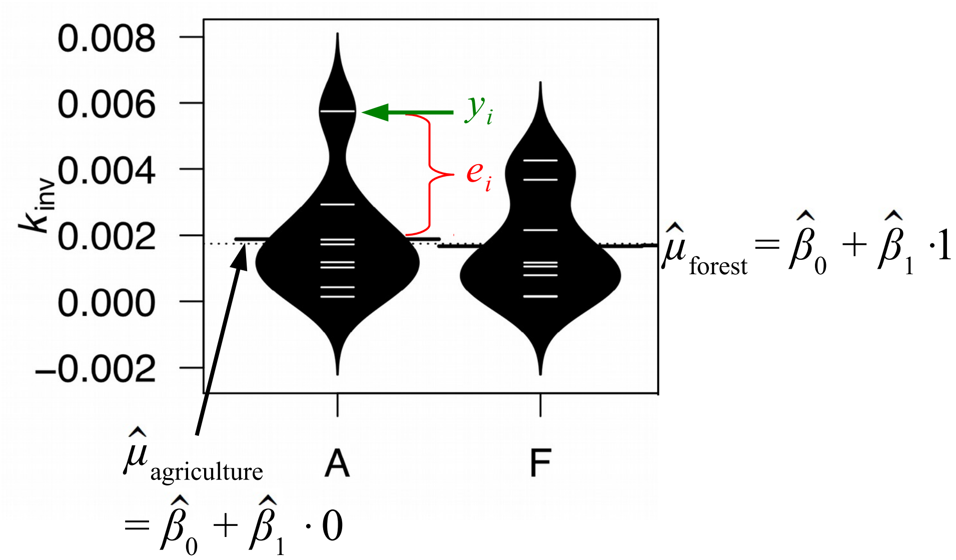
Case i	ω	$\boxed{y}$ (from R)	$\boxed{\hat{\beta}_0 + \hat{\beta}_1 X}$	$\boxed{e}$ (from R)
1	Forest	0.00016	= 0.00188 – 0.00021 +	-0.00152
2	Forest	0.00427	= 0.00188 – 0.00021 +	0.00259
3	Forest	0.00079	= 0.00188 – 0.00021 +	-0.00088
4	Forest	0.00014	= 0.00188 – 0.00021 +	-0.00154
5	Forest	0.00367	= 0.00188 – 0.00021 +	0.00200
6	Forest	0.00216	= 0.00188 – 0.00021 +	0.00048
7	Forest	0.00117	= 0.00188 – 0.00021 +	-0.00051
8	Forest	0.00105	= 0.00188 – 0.00021 +	-0.00062
9	Agriculture	0.00043	= 0.00188 + 0 +	-0.00145
10	Agriculture	0.00102	= 0.00188 + 0 +	-0.00086
11	Agriculture	0.00173	= 0.00188 + 0 +	-0.00015
12	Agriculture	0.00187	= 0.00188 + 0 +	-0.00001
13	Agriculture	0.00575	= 0.00188 + 0 +	0.00387
14	Agriculture	0.00118	= 0.00188 + 0 +	-0.00070
15	Agriculture	0.00294	= 0.00188 + 0 +	0.00105
16	Agriculture	0.00014	= 0.00188 + 0 +	-0.00175

12

b_0, b_1 represent the estimates of the true regression coefficients β_0, β_1 and e is the residual. Remember that we defined $b_0 = \hat{\beta}_0$ and $b_1 = \hat{\beta}_1$.

Slight mismatches are due to rounding.

Graphical illustration of regression with a categorical predictor



A = agriculture, F = forest

Unity of the linear model and multiple linear regression

Contents

- Categorical predictors in regression analysis
- **Analysis of variance with *F*-test and unity of the linear model**
- Linear model with multiple predictors
- Interactions
- Analysis of covariance
- Multiple inference

Categorical predictor in regression

Example with four-level predictor (land use example)

Mean Agriculture

p-values for pairwise comparisons with the Intercept (Agriculture)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0018838  0.0005660   3.328  0.00271 **
Land_typeF  -0.0002079  0.0008005  -0.260  0.79719
Land_typeU  -0.0001864  0.0008647  -0.216  0.83107
Land_typeV  -0.0001551  0.0008286  -0.187  0.85306
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001601 on 25 degrees of freedom
Multiple R-squared:  0.003219,  Adjusted R-squared:  -0.1164
F-statistic: 0.02691 on 3 and 25 DF,  p-value: 0.9939
```

Difference to intercept

Not matching (in contrast to two-level predictor)

Remember that we delayed the discussion of the interpretation of the *F*-statistic in the R output.

We will discuss in the related R session how to assess different hypotheses (i.e. how to calculate *p*-values for other comparisons). We will also discuss why the standard errors are smaller than in pairwise *t*-tests.

The pairwise comparison with the Intercept can also be interpreted as an assessment whether the respective regression coefficient is 0.

F-statistic

Remember: MSS = Model Sum of Squares
RSS = Residual Sum of Squares
MSE = Mean Squared Error

In the context of the linear regression model with p predictors (or $p+1$ levels of a factor), the statistic is used to assess H_0: $\beta_1 = \beta_2 = \ldots = \beta_p = 0$ (or H_0: $\mu_1 = \mu_2 = \ldots = \mu_{p+1}$).

The statistic is calculated as the division of two variances:

$$F_{(p,\,n-p-1)} = \frac{\left(\dfrac{\text{Explained var.}}{\boxed{\text{DoF model}}}\right)}{\left(\dfrac{\text{Unexplained var.}}{\boxed{\text{DoF error}}}\right)} = \frac{\left(\boxed{\dfrac{1}{p}}\,\text{MSS}\right)}{\left(\boxed{\dfrac{1}{n-p-1}}\,\text{RSS}\right)} = \frac{\left(\dfrac{\text{MSS}}{p}\right)}{\text{MSE}}$$

p = no. parameters (predictors) in model excluding intercept
n = sample size (no. of cases across all groups in case of factor)

16

In other words, assessing H_0 involves the comparison of the regression model without any predictor to the current model. The model without any predictor only contains the intercept, where the intercept equals the grand mean, i.e. the mean over all data (over all groups in the case of a categorical predictor).

Note that H_0 can also be expressed as Var(β) = 0 or Var(μ)= 0, respectively. This follows logically, because if all regression coefficients are zero or all means are the same, the variance of these is 0.

Also remember that:

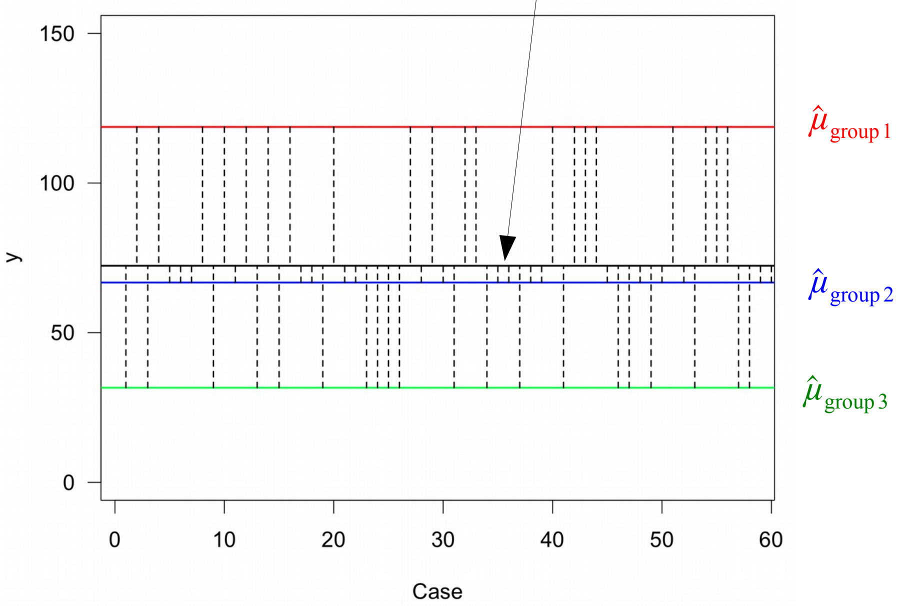
$$\text{MSE} = \frac{1}{\text{DoF}}\text{RSS} \quad \text{where DoF} = \text{Degrees of freedom}$$

We can see from the equation that the F-value increases with the MSS (explained variance) and decreases with the MSE (that is related to the unexplained variance). The related p-value decreases with an increase in the F-value. Intuitively, this makes sense because with an increase in the ratio of explained to unexplained variance, the probability decreases that the data originate from a population, for which all coefficients of the predictors are 0 or for which the mean of all factor levels is the same.

Graphical illustration of F-statistic

$$F_{(p,\,n-p-1)} = \frac{\left(\frac{\text{Explained var.}}{\text{DoF model}}\right)}{\left(\frac{\text{Unexplained var.}}{\text{DoF error}}\right)} = \frac{\left(\frac{1}{p}\boxed{\text{MSS}}\right)}{\left(\frac{1}{n-p-1}\boxed{\text{RSS}}\right)} = \frac{\left(\frac{\text{MSS}}{p}\right)}{\text{MSE}}$$

$$\boxed{\text{MSS}} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 \qquad \boxed{\text{RSS}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$



17

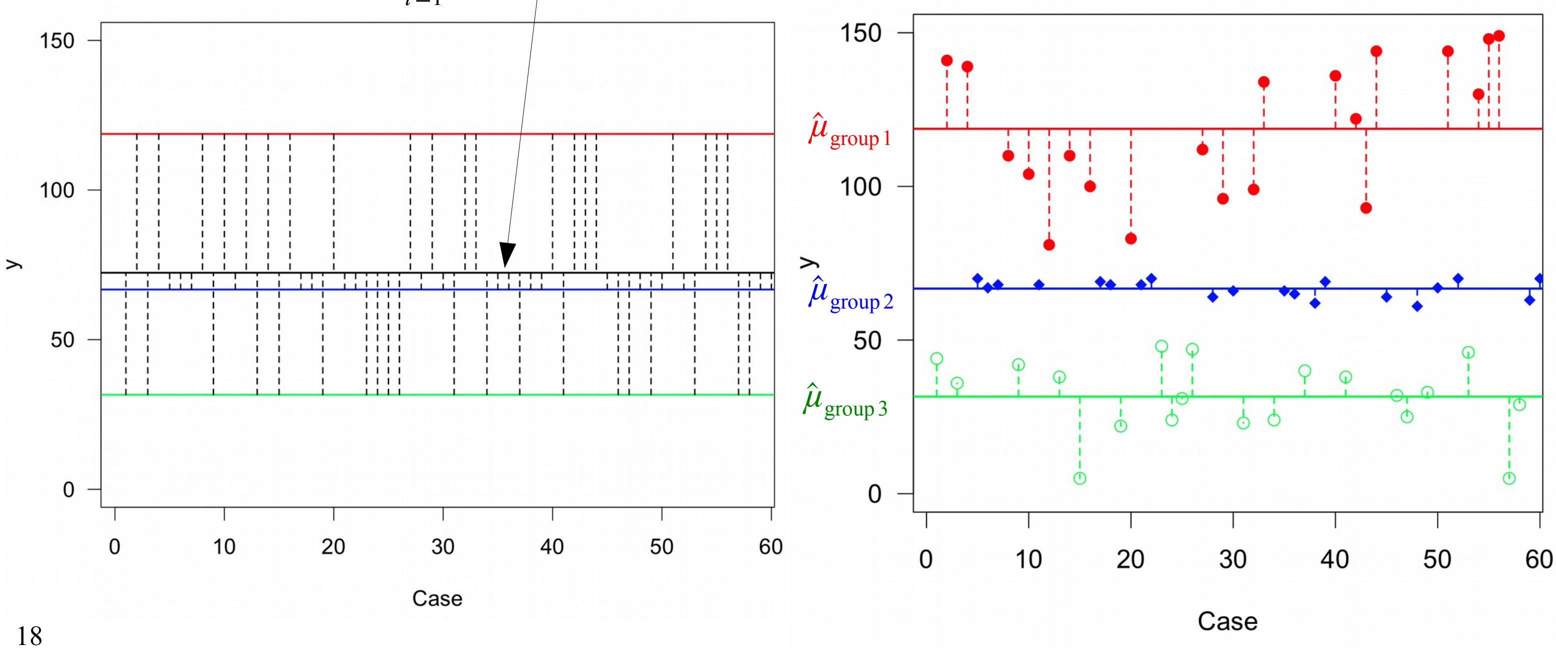
We illustrate the F-statistic for a categorical variable with j = 3 groups, i.e. 3 levels.

Remember that $\hat{y}_i$ corresponds to one of the group means $\hat{\mu}_j$.

Graphical illustration of F-statistic

$$F_{(p,\,n-p-1)} = \frac{\left(\dfrac{\text{Explained var.}}{\text{DoF model}}\right)}{\left(\dfrac{\text{Unexplained var.}}{\text{DoF error}}\right)} = \frac{\left(\dfrac{1}{p}\boxed{\text{MSS}}\right)}{\left(\dfrac{1}{n-p-1}\boxed{\text{RSS}}\right)} = \frac{\left(\dfrac{\text{MSS}}{p}\right)}{\text{MSE}}$$

$$\boxed{\text{MSS}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \qquad \boxed{\text{RSS}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$



18

We illustrate the F-statistic for a categorical variable with j = 3 groups, i.e. 3 levels.

Remember that $\hat{y}_i$ corresponds to one of the group means $\hat{\mu}_j$.

The figure also illustrates why RSS is referred to as within-group variation and MSS as between-group variation.

Analysis Of Variance (ANOVA)

- ANOVA refers to test statistics (e.g. F-statistic) related to the question: Are j sample means drawn from (statistical) populations with equal μ?
- H_0 : $\mu_1 = \mu_2 = \ldots = \mu_j$ (The related alternative hypothesis would be: $H_A: \exists i, j: \mu_i \neq \mu_j$)
- In case the ANOVA indicates different means, often multiple (pairwise) comparisons (e.g. with the t-test) are conducted afterwards to identify the means that differ. Multiple comparisons are a case of multiple inference.
- Multiple inference may require adjustment of p-values (if compared against a fixed threshold) and of confidence intervals.

19

In papers, you will often find that an ANOVA based on the F-test is simply called ANOVA (without specifying the test statistics).

We will discuss the issue of multiple inference later.

Unity of the linear model

Regression

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0018838  0.0005660   3.328  0.00271 **
Land_typeF  -0.0002079  0.0008005  -0.260  0.79719
Land_typeU  -0.0001864  0.0008647  -0.216  0.83107
Land_typeV  -0.0001551  0.0008286  -0.187  0.85306
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.001601 on 25 degrees of freedom
Multiple R-squared:  0.003219,	Adjusted R-squared:  -0.1164
F-statistic: 0.02691 on 3 and 25 DF,  p-value: 0.9939
```

ANOVA with F-test

```
          Df    Sum Sq   Mean Sq F value Pr(>F)
Land_type  3 2.100e-07 6.900e-08   0.027  0.994
Residuals 25 6.408e-05 2.563e-06
```

Square-root

20

Also note that the R^2 is obtained by dividing the sum of squares („Sum Sq“) of Land_type (in our terminology defined as Model Sum of Squares (MSS)) by the sum of the sum of squares for Land_type and Residuals (in our terminology defined as Residual Sum of Squares (RSS)). Remember that the R^2 is MSS divided by the Total Sum of Squares (TSS), where TSS = MSS + RSS. As discussed before, beware that the terminology can differ. In the context of ANOVA, MSS is frequently referred to as Treatment Sum of Squares, abbreviated TRSS, SSTR or $SS_{Treatment}$. Other terminology relates to the fact that RSS captures the within-group variation and MSS the between-group variation, and consequently uses the terms SSB or BSS and SSW or WSS for the Between Sum of Squares and Within Sum of Squares, respectively.

Note that with multiple variables, the meaning of the F-tests of the regression and ANOVA will differ.

ANOVA assumptions and alternatives

- Given unity of the linear model:
  - Assumptions of regression model apply
  - Diagnosis differs slightly (e.g. residual vs. fitted plot suboptimal) → Use methods discussed for t-test
  - Dealing with violations → Approaches discussed for regression model largely applicable
  - As for regression model, violation of variance homogeneity assumption more problematic than of normal distribution assumption
- Alternative test statistics:
  - Welch's ANOVA for data with heterogeneous variances
  - Kruskal-Wallis test or permutational ANOVA for non-normally distributed data (if variance homogeneous!)

21

The assumptions of the linear regression model (Normal distribution of error, Variance homogeneity and Independence of errors) also apply to ANOVA. The assumption of a linear relationship is obviously irrelevant for a single categorical variable, but will become relevant in models with multiple variables. Our discussion of the relevance of the violation of assumptions of the regression model also applies to ANOVA. The ANOVA (with the F-test) is more robust to violation of the normal distribution assumption than of the variance homogeneity assumption (for details see Glass et al. 1972). Contrary to suggestions in some textbooks, the Kruskal-Wallis test is no appropriate alternative in the case of heteroscedasticity (Vargha & Delaney 1998). If the data is normally distributed, Welch's ANOVA represents a reliable alternative, similar to Welch's t-test being an alternative to the classical t-test in the situation with two groups that have heterogeneous variances. In addition, and see our discussion for the regression model, non-normally distributed data could be transformed to obtain normality, but if another model such as the Generalized Linear Model (GLM) fits better to the data, then rather change the model than the data. Finally, as for the regression model, the heterogeneity of variances can be incorporated into the model using generalised least squares (see chapter 4 in Zuur et al. 2009). A robust alternative to the classical ANOVA, such as median regression to the classical regression model, has been introduced by Herberich et al. (2010). Their article features example R code that can be adapted to other data sets.

Model diagnosis, i.e. checking the assumptions, can be done before an ANOVA with one explanatory (categorical) variable. Hence, it is not necessary to fit an ANOVA for checking the assumptions. If the assumptions are not met, a different model or a different ANOVA (e.g. Welch's ANOVA) can be fitted directly. Indeed, in the case of an ANOVA with one explanatory variable, it makes no difference if the variance homogeneity across groups and the normal distribution per group is checked using the raw data or the residuals. Note that in the case of an ANOVA with multiple variables, which will be discussed later, the residuals should be used.

Many more important topics are related to ANOVA such as multiple comparisons, contrasts and mixed-effect models. They will be discussed later in the course or in the R tutorials.

References

Glass, Peckham & Sanders 1972 *Review of Educational Research* 42: 237-288; freely accessible within our university at http://www.jstor.org/pss/1169991.

Herberich E, Sikorski J, Hothorn T, 2010 A Robust Procedure for Comparing Multiple Means under Heteroscedasticity in Unbalanced Designs. PLoS ONE 5(3): e9788; freely accessible within our university at: http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0009788)

Vargha & Delaney 1998 *Journal of Educational and Behavioral Statistics* 23: 170-192; freely accessible within our university at http://www.jstor.org/pss/1165320

Zuur, A. F. et al. 2009: Mixed effects models and extensions in ecology with R. Springer: New York


LET'S USE
MULTIPLE VARIABLES!
imgflip.com

Unity of the linear model and multiple linear regression

Contents

- Categorical predictors in regression analysis
- Analysis of variance with *F*-test and unity of the linear model
- **Linear model with multiple predictors**
- Interactions
- Analysis of covariance
- Multiple inference

Linear model with multiple predictors

- Used to establish a (linear) relationship between multiple explanatory (predictor) and one response variable
- Model types:
  - Multiple continuous variables → Multiple linear regression
  - Multiple categorical variables → Multi-way ANOVA
  - Mix of continuous and categorical variables → ANCOVA
- Research goals same as for single-predictor linear model:
  - Prediction (predict response to multiple predictors)
  - Estimation (of regression coefficients and explained variance)
  - Assessing hypotheses (regarding relationship of response with multiple predictors)
  - Explanation (identify important explanatory variables)

Compared to the case with one explanatory variable or predictor, the research goals remain fairly similar. Regarding prediction, the goal is to predict a response based on a linear relationship of this response with multiple predictors. Similarly, the goal estimation is largely the same in the context of multiple variables as described before. Imagine that instead of asking “How much does plant growth increase per additional unit of nutrients?”, we ask: “How much does plant growth increase per additional unit of nutrients, temperature and irrigation?”. The estimation goal in multiple linear regression and in ANCOVA typically relates to the regression coefficients, whereas in multi-way ANOVA it typically relates to the fraction of variance explained by different explanatory variables.

The research goal explanation will be discussed with an example on the next slide.

In fact, you can use the term multiple linear regression for all types of models, i.e. it is rather a hypernym. Nevertheless, if you have specific questions regarding a model with multiple categorical predictors or how categorical predictors influence the interpretation of the slopes of continuous predictors in the cased of mixed predictors, you are typically better off screening textbooks for multi-way ANOVA or ANCOVA, respectively. Although textbook sections on multiple linear regression may also explain related questions, they typically focus predominantly on continuous predictors.

Linear model with multiple predictors

Research goal: Explanation (identify important explanatory variables)

<u>Example:</u> Which variable(s) do best explain the response of different groups of organisms?

Table 2. Environmental Variables Selected in Linear Model Building with Highest Explanatory Power for the Response Variables Using Explained Variance (r^2) and the Akaike Information Criterion (AIC) as Goodness of Fit Measures

response variable	log mTU_{DM}	T (°C)	conductivity (μS/cm)	turbidity (NTU)	r^2	AIC
$SPEAR_{pesticides}$	x				0.67	−34
SIGNAL	x				0.36	98
bacteria[a]						
flagellates[a]		x	x		0.49	434
ciliates[a]		x		x	0.59	209
amoebas[a]				x	0.78	200

Schäfer et al. 2011

As stated before, the goal of explanation becomes much more relevant with multiple explanatory variables that, in many cases, offer competing explanations for the response. Hence, the goal is to identify the explanatory variable or variables with the highest explanatory power.

In the example, we measured several stressors that may explain an ecological index or the total abundance of an organism group, and our research goal was to identify the variable(s) with the highest explanatory power for the index. The study was conducted in 24 streams in South-East Australia and was to designed to detect potential relationships between pesticides and invertebrates as well as microorganisms. The SPEAR index has been developed to indicate pesticide effects in invertebrate communities, whereas the SIGNAL index has been developed to indicate general ecological degradation. The log mTU is a proxy for the pesticide toxicity in a sampling site, calculated for pesticides in 24 South-East Australian streams.

Schäfer R.B., Pettigrove V., Rose G., Allinson G., Wightwick A., von der Ohe P.C., et al. (2011) Effects of pesticides monitored with three sampling methods in 24 sites on macroinvertebrates and microorganisms. Environmental Science & Technology 45, 1665–1672.

Multiple linear regression model

- Extension of simple linear regression model, we assume true relationship is: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$

  → Classical definition for case i:

  $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + \varepsilon_i$ with $\varepsilon \sim \text{Normal}(0, \sigma)$

- Using sample date, we estimate β's (b's = regression coefficients) to obtain estimates for y:

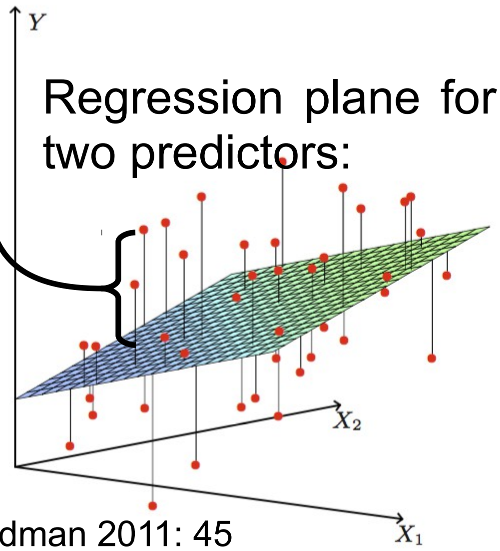
  $\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}$

- Remember: Residual e_i defined as:

  $$e_i = y_i - \hat{y}_i$$

- Model fitting through minimising the squared sum of residuals (RSS):

$$\text{Find} \arg \min_{b_0, b_1, b_2, \ldots, b_p} \sum_{i=1}^{n} \left(y_i - \left(b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p} \right) \right)^2$$



Taken from Hastie, Tibshirani and Friedman 2011: 45

If you are not familiar with the notation, revisit the material of the second session, where we introduced the simple linear regression model. We only discuss the classical definition here, for the probability distribution centric definition please also refer to the second session.

Multiple linear regression model

Model in matrix form

$$\hat{y}_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_p x_{i,p}$$

Notation for the number of observations (*n*):

$$\begin{aligned} \hat{y}_1 &= b_0 + b_1 x_{1,1} + b_2 x_{1,2} + \ldots + b_p x_{1,p} \\ \hat{y}_2 &= b_0 + b_1 x_{2,1} + b_2 x_{2,2} + \ldots + b_p x_{2,p} \\ &\vdots \\ \hat{y}_n &= b_0 + b_1 x_{n,1} + b_2 x_{n,2} + \ldots + b_p x_{n,p} \end{aligned} \quad \xrightarrow{\text{matrix}} \quad \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$$

$$\hat{\boldsymbol{Y}} = \boldsymbol{X} b$$

Matrix calculations of regression coefficients same as explained for simple linear regression model

Interpretation of individual coefficients: Provide the effect size for one unit increase in the predictor on the response, if all other predictors remain constant.

27

Note that the interpretation of coefficients under the condition that the other predictors remain constant, is problematic if the variables are correlated, i.e. the increase in one variable typically co-occurs with an increase in another variable. This issue will be discussed later in the context of collinearity in multiple linear regression analysis.

Multi-way ANOVA

- Two-way ANOVA = Two categorical predictors, Three-way ANOVA = Three categorical predictors and so on
- Example: Effect of sex and population on possum length

Regression output

```
Call:
lm(formula = totlngth ~ sex * Pop, data = pos_dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.3333  -2.5227   0.5544   2.9773   9.4949 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  87.3684     0.9880  88.426   <2e-16 ***
sexm         -0.8633     1.2049  -0.716    0.475    
PopVic        0.9649     1.3225   0.730    0.467    
sexm:PopVic  -0.9473     1.7515  -0.541    0.590    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.307 on 100 degrees of freedom
Multiple R-squared:  0.03083,	Adjusted R-squared:  0.00175 
F-statistic:  1.06 on 3 and 100 DF,  p-value: 0.3696
```

Mean of females in the other population

p-values for pairwise comparisons with Intercept (hypothesis: coefficient = 0)

Interaction term – discussed later

Square-root

ANOVA output

```
           Df Sum Sq Mean Sq F value Pr(>F)
sex         1   49.1   49.12   2.648  0.107
Pop         1    4.5    4.45   0.240  0.625
sex:Pop     1    5.4    5.43   0.293  0.590
Residuals 100 1854.8   18.55
```

3

F-statistics differ!

28

Both factors have two levels (sex: m (male) and f (female); Pop: Vic (Victoria) and other).

The F-tests differ. In the case of regression, the F-statistic relates to the comparison of the given regression model with all variables to the model that only contains an intercept (equivalent to the overall mean), also called *null model*. In the case of ANOVA, the F-statistic is calculated differently, which we will discuss hereafter.

Variance partitioning in multi-way ANOVA

lm(formula = totlngth ~ sex + Pop, data = pos_dat)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sex	1	49.1	49.12	2.648	0.107
Pop	1	4.5	4.45	0.240	0.625
sex:Pop	1	5.4	5.43	0.293	0.590
Residuals	100	1854.8	18.55		

lm(formula = totlngth ~ Pop + sex, data = pos_dat)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Pop	1	11.8	11.84	0.639	0.426
sex	1	41.7	41.72	2.250	0.137
Pop:sex	1	5.4	5.43	0.293	0.590
Residuals	100	1854.8	18.55		

29

The issue of the type of ANOVA is often irrelevant, if the ANOVA is balanced. “Balanced” means that each factor level has the same number of observations.

Variance partitioning in multi-way ANOVA

lm(formula = totlngth ~ sex + Pop, data = pos_dat)

SS(sex)
SS(Pop | sex)
SS(sex × Pop | sex, Pop)

```
          Df Sum Sq Mean Sq F value Pr(>F)
sex        1   49.1   49.12   2.648  0.107
Pop        1    4.5    4.45   0.240  0.625
sex:Pop    1    5.4    5.43   0.293  0.590
Residuals 100 1854.8  18.55
```

lm(formula = totlngth ~ Pop + sex, data = pos_dat)

SS(Pop)
SS(sex | Pop)
SS(sex × Pop | Pop, sex)

```
          Df Sum Sq Mean Sq F value Pr(>F)
Pop        1   11.8   11.84   0.639  0.426
sex        1   41.7   41.72   2.250  0.137
Pop:sex    1    5.4    5.43   0.293  0.590
Residuals 100 1854.8  18.55
```

- Difference due to Type 1 ANOVA, i.e. sequential SS, for unbalanced design
- Type 2: SS(sex | Pop) and SS(Pop | sex) as in Type 1 (→ sequence of model terms irrelevant), ignores interactions
- Type 3 considers interactions (e.g. SS (sex | Pop, sex × Pop))
- If aim is to assess hypotheses or partitioning of variance and interaction is irrelevant, use Type 2, otherwise Type 3.

30

The issue of the type of ANOVA is irrelevant if the ANOVA is balanced. “Balanced” means that each factor level, and combination of factor levels, has the same number of observations (see Table below). This is, because in a balanced design, the correlation between all factor levels is 0 and therefore each factor explains a unique fraction of the variance (i.e. sum of squares) of the response.

Table: Example for balanced design

	Sex: Male	Sex: Female
Population: Victoria	10	10
Population: Other	10	10

If you are not familiar with the Sum of Squares concept, revisit the first part of this session.

For further details on the different types of ANOVA and how to fit related models see Fox & Weisberg (2019): 260-267.

Unity of the linear model and multiple linear regression

Contents

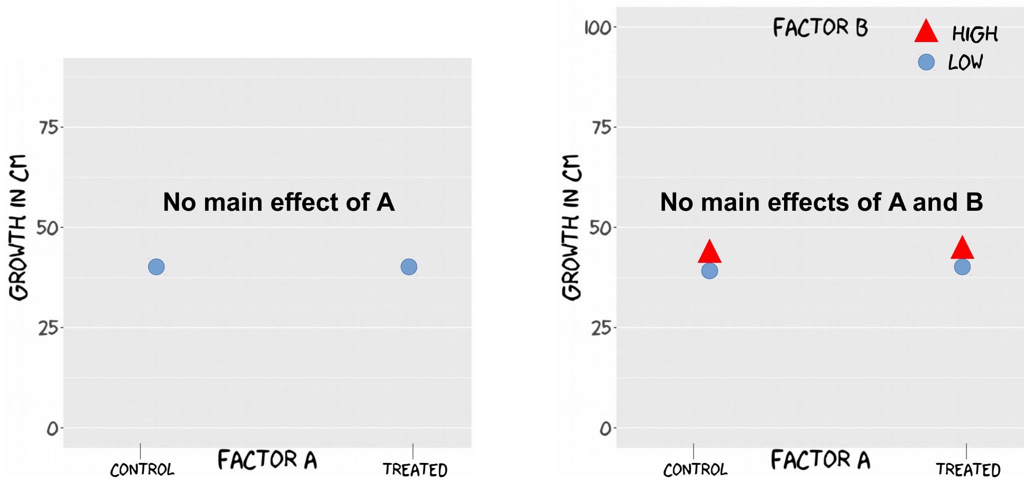
- Categorical predictors in regression analysis
- Analysis of variance with *F*-test and unity of the linear model
- Linear model with multiple predictors
- **Interactions**
- Analysis of covariance
- Multiple inference

Interactions

- Definition for linear model: Effect of one predictor on the response depends on state of another predictor
- Notation for two variables: $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1} x_{i,2}$
  Common short notation: $A + B + A \times B$

Main effects Interaction effect

Illustration:

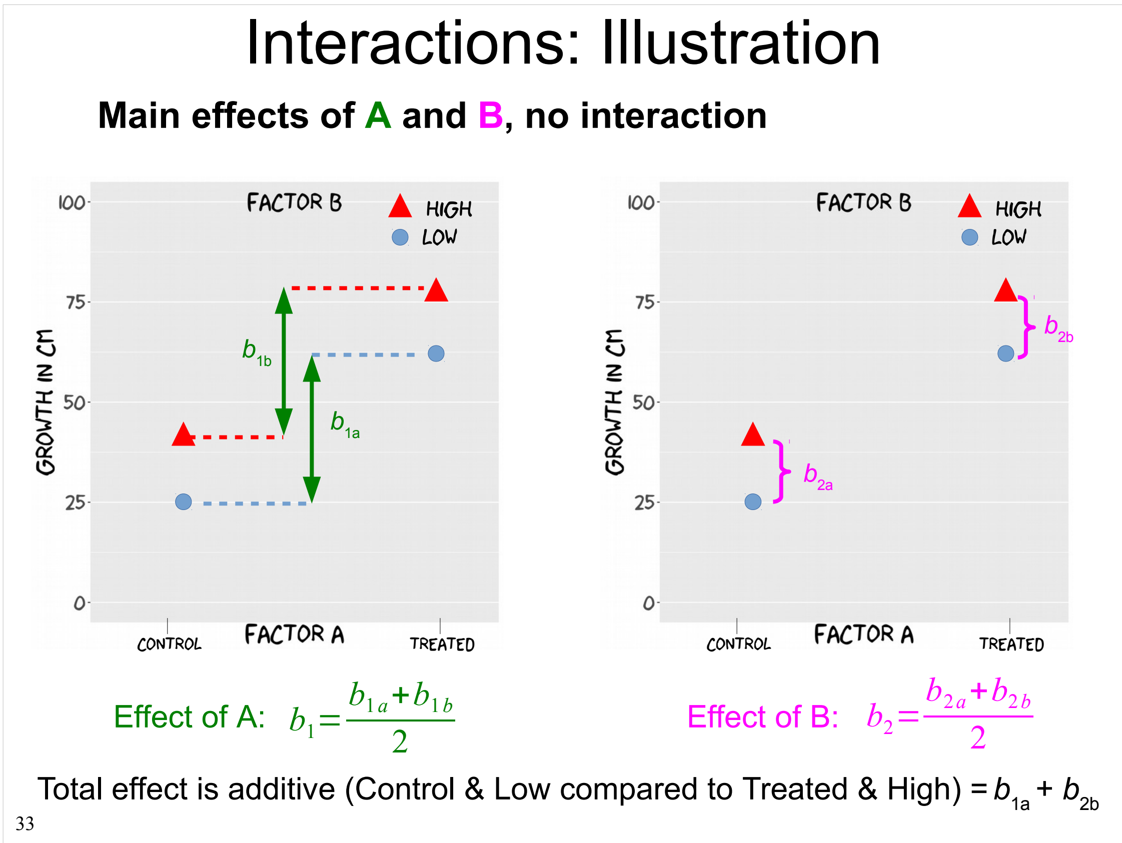


32

We discuss the topic of interaction with respect to two variables. An interaction can, of course, also occur for multiple predictor variables, which will increase the complexity of the interpretation. For example, the sex of a species could influence its sensitivity to a chemical, and a third variable, climate, could moderate these relationships.

The definition applies to both ANOVA and regression. In the case of ANOVA, an interaction implies that the effect of one factor on the response depends on the levels of a second factor (or multiple other factors). In the case of regression, an interaction implies that the relationship between a continuous predictor and the response, which is given by the slope, is moderated by other continuous variables or factors (the latter will be discussed soon in the context of analysis of covariance (ANCOVA)).

For sake of simplicity, we only plot single symbols instead of means with standard deviation.

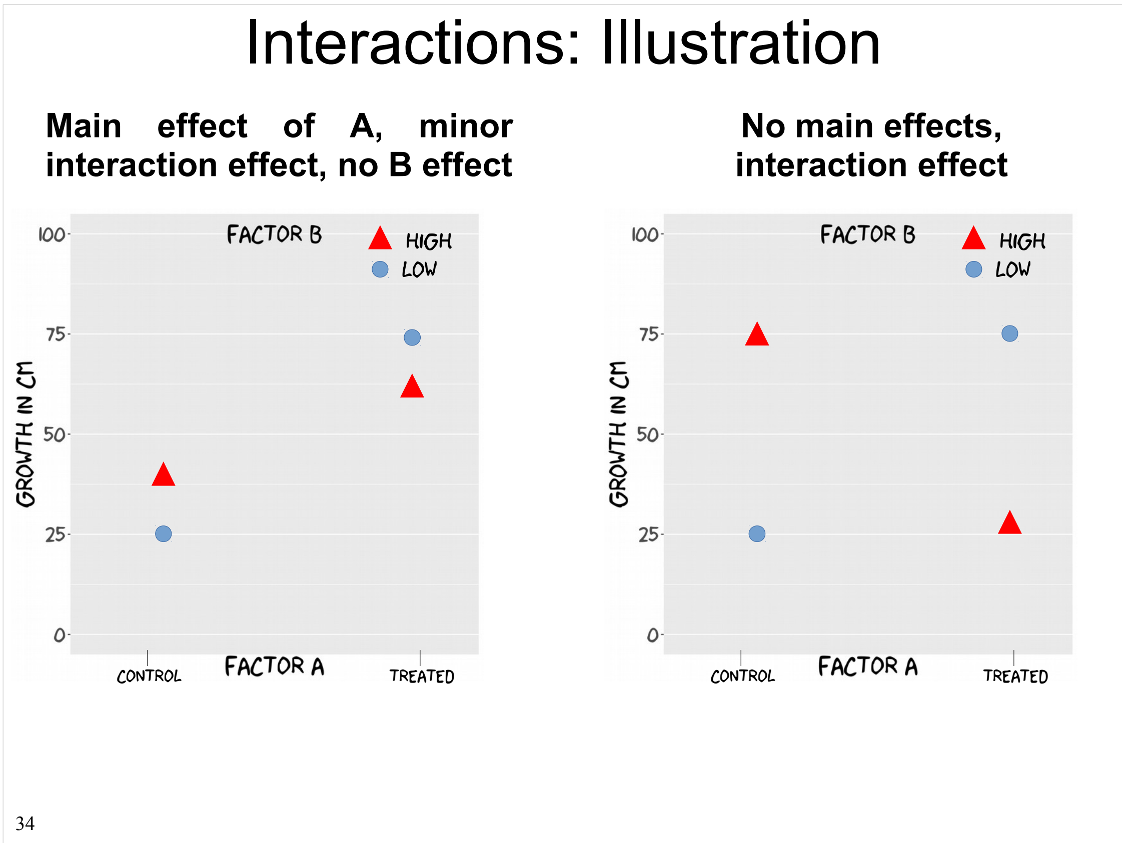


In this example, the joint effect of two treatments is additive. How the individual effects of two treatments that are stressors add up, is a key question in the context of multiple stress research (for an overview see Schäfer & Piggott 2018).

Note that additivity in a model with a log transformed response variable implies non-additivity (i.e. a multiplicative effect) on the non-transformed scale, which several researchers have been unaware of (Griffin et al. 2016).

Griffen B., Belgrad B., Cannizzo Z., Knotts E. & Hancock E. (2016) Rethinking our approach to multiple stressor studies in marine environments. Marine Ecology Progress Series 543, 273–281.

Schäfer R.B. & Piggott J.J. (2018) Advancing understanding and prediction in multiple stressor research through a mechanistic basis for null models. Global Change Biology 24, 1817–1826.

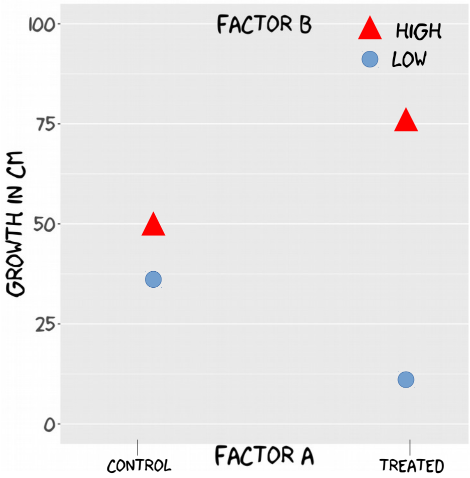


In the left plot, there is no effect of B because the difference between the levels low and high of Factor B are positive and negative, respectively, and of similar size, i.e. the average effect is 0. In other words, the sum of b_{2a} and b_{2b} is 0 (see previous slide for equation).

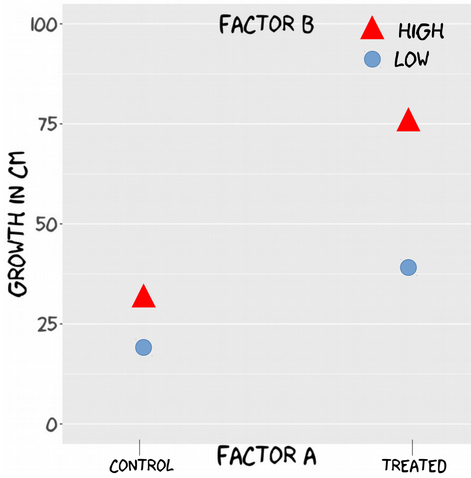
The same applies to the right plot. In this plot, the effect of A is also 0 for the same reason, i.e. b_{1a} + b_{1b} = 0 given that the increase of low from control to treatment is similar to the decrease of high from control to treatment.

Interactions: Illustration

Main effect of B, large interaction effect, no A effect



Main effect of A and B, interaction effect

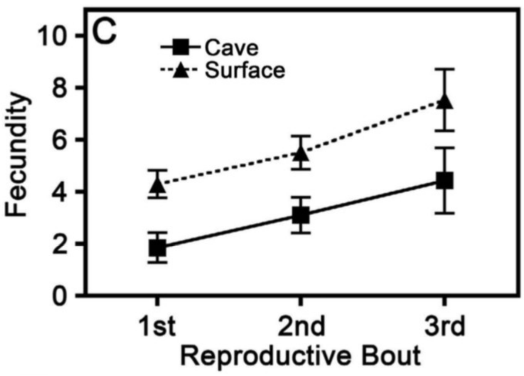


35

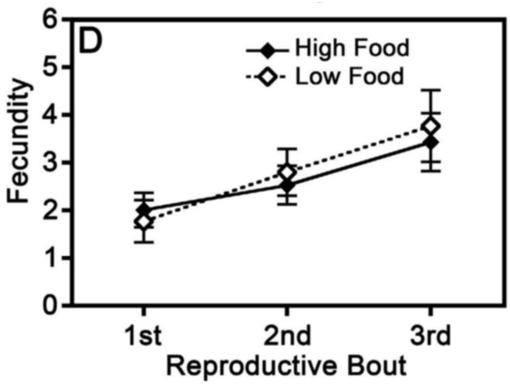
Interactions: Real world example

Study on the adaptation of mollis from different origins (caves or surface water)

Main effect of origin and reproductive bout, no interaction



Main effect of reproduction, no interaction effect and of food

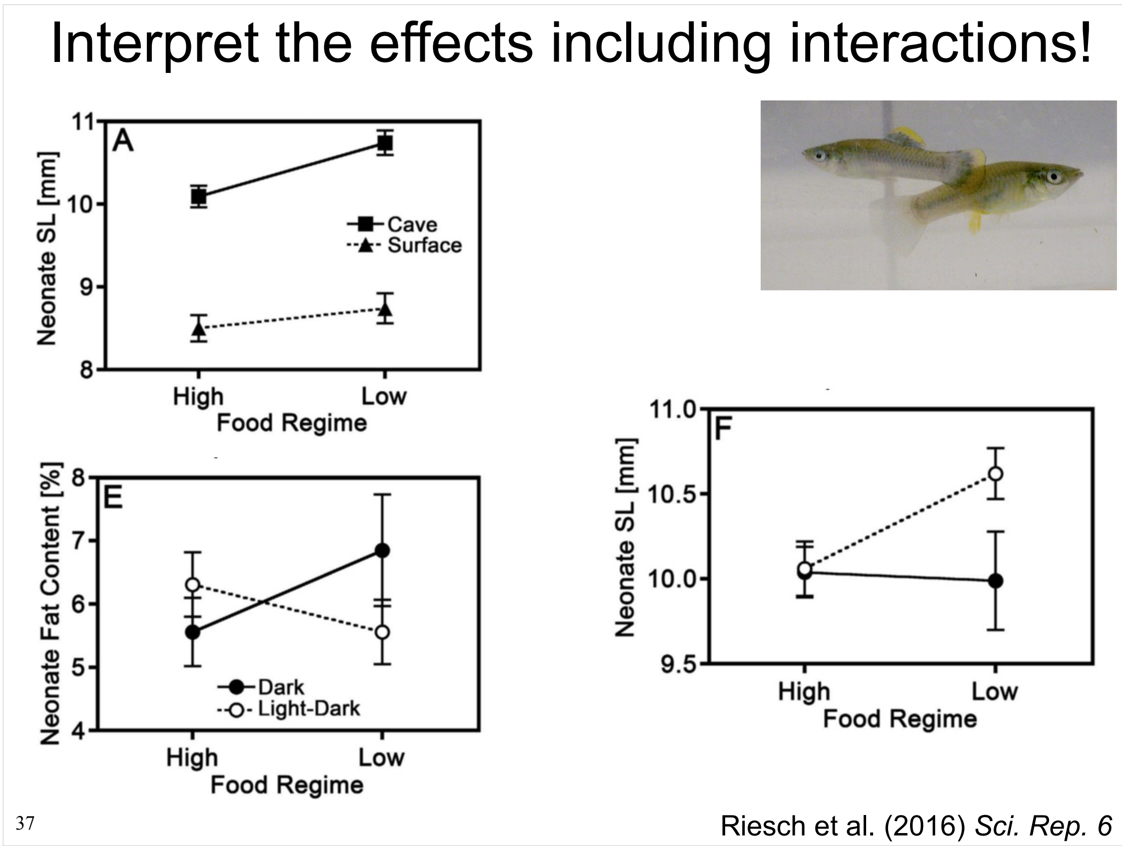


36

Riesch et al. (2016) *Sci. Rep. 6*

Picture taken from:
https://en.wikipedia.org/wiki/Poecilia_mexicana#/media/File:Poecilia_mexcana.JPG

Riesch R., Reznick D.N., Plath M. & Schlupp I. (2016) Sex-specific local life-history adaptation in surface- and cave-dwelling Atlantic mollies (Poecilia mexicana). *Scientific Reports 6*.



Picture taken from: https://en.wikipedia.org/wiki/Poecilia_mexicana#/media/File:Poecilia_mexcana.JPG

Figure A shows an effect of food (average increase from high to low), an effect of the origin (average increase from surface to cave) and a negligible interaction (increase from high to low food slightly higher for cave).

Figure E displays minor effects of the two main factors, but a clear interaction.

Figure F presents a clear interaction effect: No change in the effect of dark light regime from high to low food, but in the effect of the light-dark regime.

Overall, both E and F provide examples of the difficulty to interpret the main effects in the presence of interactions.

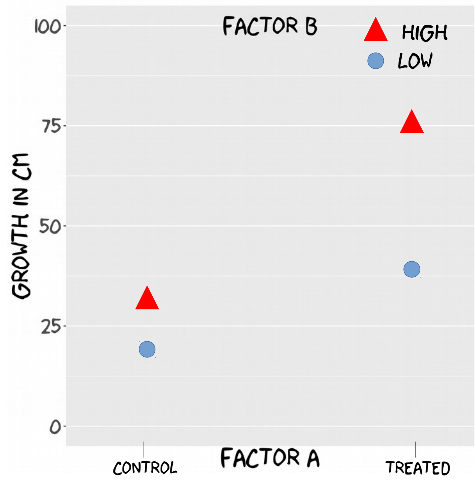
Riesch R., Reznick D.N., Plath M. & Schlupp I. (2016) Sex-specific local life-history adaptation in surface- and cave-dwelling Atlantic mollies (Poecilia mexicana). *Scientific Reports 6*.

Interpreting main effects in presence of interactions

When is an interaction present?

- Classic approach: p-value < 0.05
- Visual inspection

Illustration:



Simplified, idealized R output

ANOVA (summary.aov)

	Pr(>F)
Factor 1	<0.001
Factor 2	<0.001
Factor1:Factor2	<0.001

Main effect of A and B can be interpreted as increasing growth, despite interaction effect

38

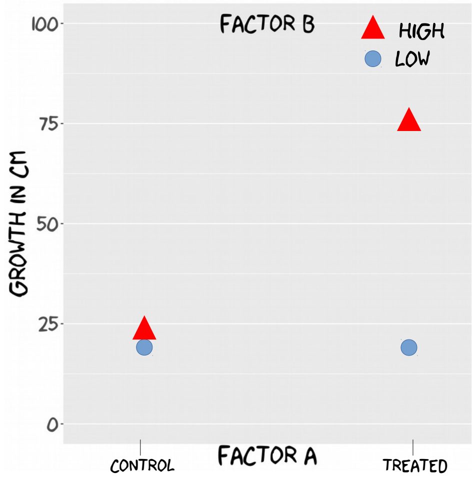
We do not delve too much into the discussion of using the p-value to assess the presence of an interaction. However, note that all the caveats regarding the dichotomisation of the p-value scale that we discussed in the last session apply here as well.

The example on the slide provides a case where the main effects can be interpreted despite the presence of an interaction.

The design is balanced, hence we do not need to worry about the type of ANOVA.

Interpreting main effects in presence of interactions

Illustration:



Simplified, idealized R output

ANOVA (summary.aov)

	Pr(>F)
Factor 1	<0.001
Factor 2	<0.001
Factor1:Factor2	<0.001

Interpretation of main effects of A and B is misleading

39

The design is balanced, hence we do not need to worry about the type of ANOVA.

Final notes on interpretation of main effects in presence of interactions

- If Sum of Squares and coefficients of interaction term(s) << main terms, interaction likely irrelevant
- Interpret only after visual checking
- Multiple linear regression: Standardize data before including interaction terms:
  - Reduces collinearity with main variables
  - Improves interpretation

40

If an interaction term is present in a multiple linear regression model, the slope of the main variable can only be taken at face value for x = 0 of the variable(s) involved in the interaction. When standardizing variables, the data is centred to the mean (mean is subtracted from each observation) and divided by the standard deviation. This means that for the standardized variables x = 0 will lie approximately in the centre of the data range. Hence, the interpretation of the slope, when the other variables are at x = 0, becomes more meaningful for standardized data. You can find more detailed information in Quinn & Keough (2002), p. 130-131 and in Fox & Weisberg (2019): 222-223.

Harrell (2015) and Fox (2015, in particular p. 160-164) deal with interactions at many places throughout their books and provide guidance on the calculation and interpretation of interactions. Underwood (1997), p. 318-323, gives a very readable introduction on interpreting interactions in the context of two-way ANOVA based on an ecological example. Finally, Fox and Weisberg (2019) outline how interactions can be included in models and examined in R.

Underwood A.J. (1997) Experiments in ecology: their logical design and interpretation using analysis of variance. Cambridge University Press, New York, NY, USA.

Unity of the linear model and multiple linear regression

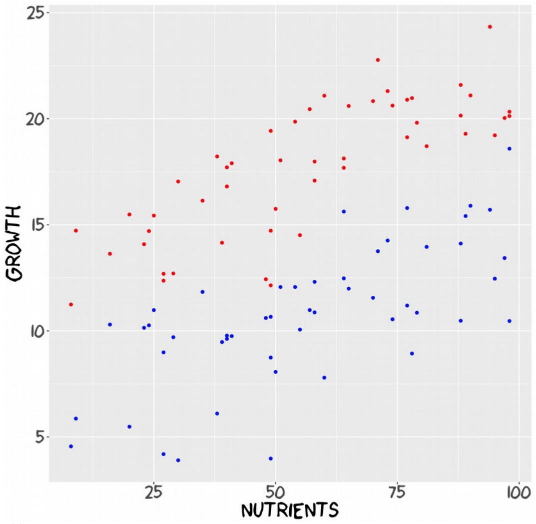
<u>Contents</u>

- Categorical predictors in regression analysis
- Analysis of variance with *F*-test and unity of the linear model
- Linear model with multiple predictors
- Interactions
- **Analysis of covariance**
- Multiple inference

Analysis Of Covariance (ANCOVA)

- Combination of continuous and categorical predictors

<u>Experiment: Response of plant growth to nutrients in two soils</u>



We could fit data to two independent regression models:

$$Y_{\text{Clay}} = \beta_{0,\text{ Clay}} + \beta_{1,\text{ Clay}} X_1 + \varepsilon$$

$$Y_{\text{Loam}} = \beta_{0,\text{ Loam}} + \beta_{1,\text{ Loam}} X_1 + \varepsilon$$

But what about a single model?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

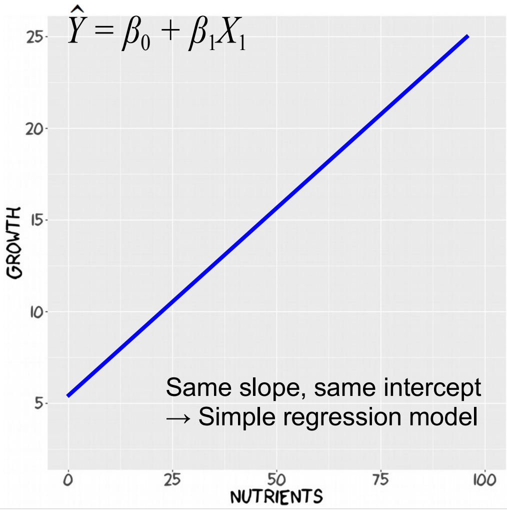
$$X_2(\omega) = \begin{Bmatrix} 0 & = & \text{Loam} \\ 1 & = & \text{Clay} \end{Bmatrix}$$

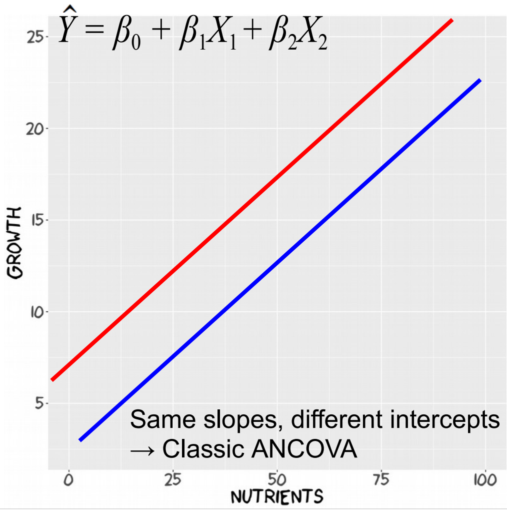
42

Quantitative, continuous explanatory variables/ predictors are also called covariates.

Analysis Of Covariance (ANCOVA)

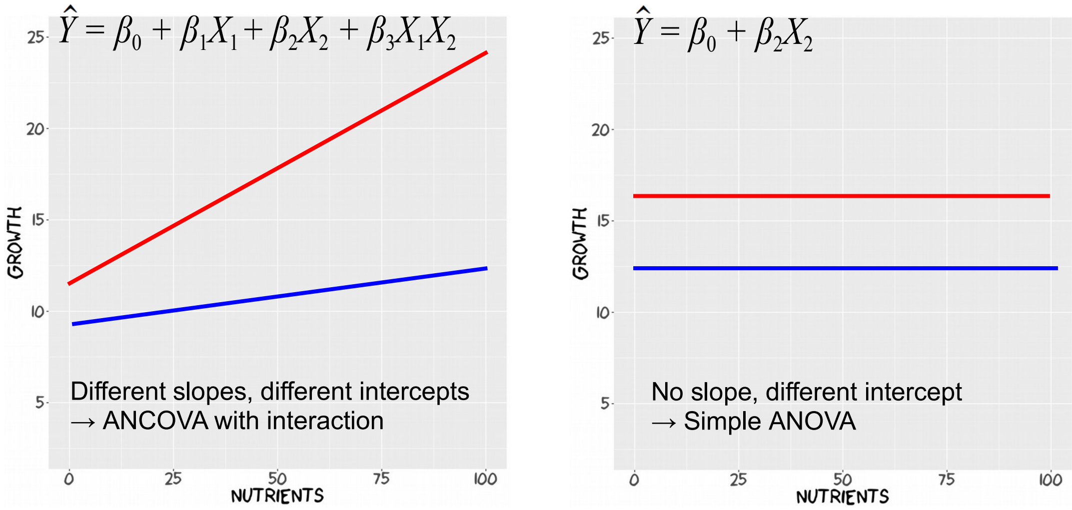
- Advantage of single model:
  - Less parameters compared to independent models
  - More comprehensive interpretation
  - Allows for assessment of different hypotheses (e.g. slopes are equal, intercept equal)





43

Analysis Of Covariance (ANCOVA)

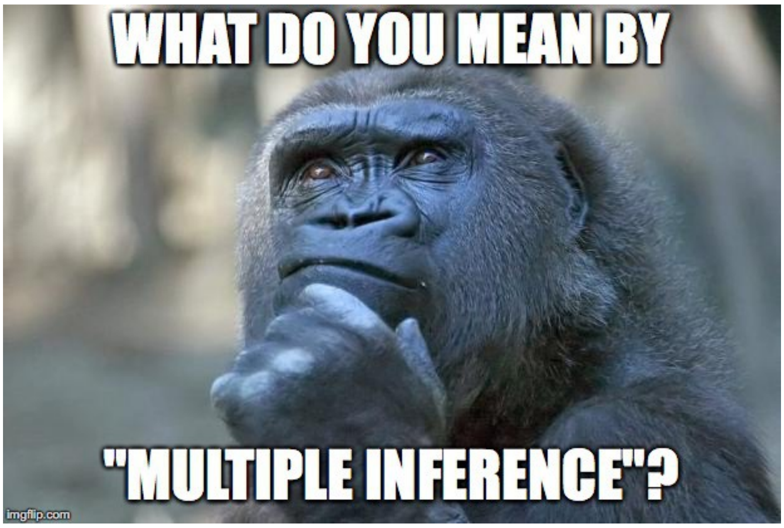


- How to decide? Sequential assessment of hypotheses: 1. $\beta_3 = 0$? Yes → 2. $\beta_2 = 0$ or $\beta_1 = 0$?
- Issue of multiple inference

44

Note that if the assessment of the first hypothesis concludes that there is no interaction, the interaction term should be removed from the model as otherwise the interpretation may be flawed. For details see Engqvist (2005).

Engqvist L. (2005) The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. Animal Behaviour 70, 967–971.


WHAT DO YOU MEAN BY
"MULTIPLE INFERENCE"?
imgflip.com

Unity of the linear model and multiple linear regression

Contents

Multiple inference: CIs

Background: If computing 100 CIs, 5 will not include the true parameter. If we will calculate several CIs from future study data what is the probability that one of them excludes the true value?

1 CI: $1-(0.95) = 0.05$

2 CIs: $1-(0.95)^2 = 0.1$

…

5 CIs: $1-(0.95)^5 = 0.23$

If the objective is to keep the probability across all CIs at 0.05 (called family-wise error rate), you need to adjust the confidence level.

Example for adjustment: 5 CIs: $1-(0.99)^5 = 0.05$

47

The case study serves as illustration for the rationale of multiple inference. You may be familiar with other examples, where the probability of the occurrence of an outcome of an event (i.e. a draw from a random variable) with small probability becomes relatively high in the case of multiple events. For example, although the probability of having the same birthday as another person is just 1/365, already for 23 persons the chance is 50% that two persons have birthday on the same day. Generally, even if independent events have a small probability of occurrence, there can be a substantial probability that one of them occurs (see Matloff 2017: 295f).

The adjustment in the given examples means that the confidence level for each single CI is adjusted to 0.99 (from 0.95) to control the overall (so-called familywise) confidence level.

Multiple inference: p-values

- Same issue when assessing hypotheses using a fixed significance threshold (i.e. following PaleoFisherian approach using a significance level α = 0.05)
- When do I have to adjust?
  - Difference between selective and simultaneous inference.
    - Selective inference: Inference on a subset of data that has been manually or automatically selected in the light of data
    - Simultaneous inference: Inference on all possible relationships or comparisons in the data set
  - In case of selective inference and in case of simultaneous inference, if parts of data re-used in statistical procedure
  - Not in case of simultaneous inference, if statistical procedures (e.g. specific hypotheses for assessment) and data pre-processing have been defined *a priori* and data (re-)used is independent (e.g. orthogonal contrasts)
  - Not consistently handled in scientific literature

48

The multiple inference issue is independent from the size of the fixed threshold, it would be the same for any other (plausible) significance level.

The difference between simultaneous and selective inference is nicely described in Cox (1965) and with examples for more recent approaches of data analysis in Taylor and Tibshirani (2015).

Data are re-used, if the same variable is, for example, used in several correlation analyses. By contrast, if we decided on the data pre-processing before exploring the data (see the topic of researcher degrees of freedom in Gelman 2013) and defined before the study to assess two specific hypotheses that do not involve the re-use of a variable/data, e.g. correlation of *a* with *b* and of *c* with *d*, no adjustment for multiple inference is required. However, in specific cases such as orthogonal contrasts, the re-use of data does not imply adjustment for multiple inference.

Contrasts will be further explained in the R demonstration. For details on contrasts see Fox and Weisberger (2019), Bretz et al. 2011 and Crawley (2012). You find a simple introduction into contrasts in the literature folder on the OpenOLAT course page.

Though it is clear that many statistical tests with the same data (ironically called “fishing expedition” - for example, when you calculate all possible correlations between variables with the aim to find one with a low p-value) should be punished with lower significance thresholds, many cases are less obvious and not always handled consistently in the literature. For example, in the regression context, reviewers typically do not ask to adjust for conducting multiple t-tests, whereas this is often requested in the ANOVA context. To be consistent, inference on the same data that is used in multiple studies should be adjusted, which is rarely if ever done. Note that in general the choice when to adjust is somehow arbitrary. Instead of adjusting for multiple inference for a single study, we could also adjust across studies of the same author(s), across all studies in an issue of a journal etc.

Bretz F., Hothorn T. & Westfall P.H. (2011) Multiple comparisons using R. CRC Press, Boca Raton, FL.

Cox D.R. (1965) A Remark on Multiple Comparison Methods. Technometrics 7, 223.

Gelman A. & Loken E. (2013) The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Taylor J. & Tibshirani R.J. (2015) Statistical learning and selective inference. Proceedings of the National Academy of Sciences 112, 7629–7634.

Multiple inference: How to adjust?

- How to adjust if I have to adjust?
  - Many approaches for adjusting p-values (for details and implementation in R see Bretz et al. 2011)
  - Classical approach: Correct for number of statistical procedures (e.g. tests) using, for example, Bonferroni correction (focus is on family-wise error rate)
  - More recent approach: Control the false discovery rate (rate of falsely rejected null hypotheses). Involves the stepwise adjustment of p-values (or CIs)
  - Alternative: Move from hypothesis testing (particularly in the Neyman-Pearson or PaleoFisherian paradigm) to parameter estimation (e.g. with Bayesian models (Gelman et al. 2012))

49

We do not discuss solutions to the issue of multiple inference in detail. You find detailed guidance in Bretz et al. (2011) and in Fox and Weisberger (2019).

The false discovery rate approach has been hugely influential, the original paper by Benjamini & Hochberg (1995) is the most cited in the area of "multiple comparisons" and has received ~35,000 citations (October 2018). The second most cited paper in the area is 40 years older and has received ~5,000 citations. An overview of the approach and subsequent developments is provided in Benjamini (2010) and, with a special focus on ecology, in Pike (2011).

The Bonferroni correction treats each *p*-value the same, whereas the method of Benjamini & Hochberg (1995) works sequentially on *p*-values. They are sorted from lowest to highest and subsequently low values are stronger corrected than high values.

Benjamini Y. & Hochberg Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B-Methodological 57, 289–300.

Benjamini Y. (2010) Discovering the false discovery rate: False Discovery Rate. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 405–416.

Bretz F., Hothorn T. & Westfall P.H. (2011) Multiple comparisons using R. CRC Press, Boca Raton, FL.

Gelman A., Hill J. & Yajima M. (2012) Why We (Usually) Don't Have to Worry About Multiple Comparisons. Journal of Research on Educational Effectiveness 5, 189–211.

Pike N. (2011) Using false discovery rates for multiple comparisons in ecology and evolution: False discovery rates for multiple comparisons. Methods in Ecology and Evolution 2, 278–282.