# Statistik III

# Advanced Statistics

Florian Wickelmaier

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN



Summer 2026

# General Linear Model

Basics

- The general linear model is an abstract statistical model class that relates quantitative (continuous) and/or qualitative (categorical) predictor variables to one or more response variables
- It includes a variety of familiar statistical procedures as special cases
  - Regression
  - Analysis of variance
  - Analysis of covariance
- It offers a formal framework for parameter estimation, hypothesis testing, and implementation in software (e. g., R, SPSS, SAS)

# General Linear Model

Assumptions

- For each data point the expected value of the response variables is computed in a simple way as a linear combination of the values of the predictor variables
- Assumptions regarding error (i. e., the deviation of the actual values of the response variable from the respective expected values)
  - Stochastic independence from the response variable
  - Variance homogeneity
  - Normal distribution

# General Linear Model

Basics

- In the sequel all effects (except for the error) will be understood as *fixed effects*
- *Random effects* will be introduced later (mixed models)
- To begin with, the following two-way analysis of variance is confined to a factorial and balanced design
  - Equal number of observations in all of the factor level combinations
- Unbalanced designs and their consequences for hypotheses testing are considered in a second step

# General Linear Model

Model equation

- The model equation is structurally identical to multiple regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Notation

| | |
|---|---|
| $\mathbf{y}$ | $n$ random vector of the response variable |
| $\mathbf{X}$ | $(n \times m)$ design matrix (not necessarily a random matrix) |
| $\boldsymbol{\beta}$ | $m$ parameter vector |
| $\boldsymbol{\varepsilon}$ | $n$ random vector representing error |

# General Linear Model

Model equation

- The $(n \times m)$ design matrix $\mathbf{X} = (x_{ij})$ specifies the values of the independent variables
  - The number of rows $n$ equals the number of observations
  - The number of columns $m$ is in general not identical to the number of predictor variables, but is indirectly related to it (see below for examples)
- Categorical variables (such as conditions in ANOVA) are to be "represented" appropriately in the design matrix $\mathbf{X}$ (via some coding scheme)

# General Linear Model

Model equation

- In general, the parameter vector $\boldsymbol{\beta}$ describes the impact that the predictor variables exert on the response variable
- For example, $\beta_j$ characterizes the effect of the variable

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

- The interpretation of $\beta_j$ critically depends on the coding of the predictor variable in the design matrix $\mathbf{X}$, as illustrated in the following applications of the general linear model

# Multiple Regression

Model equation

- Prediction of $Y$ through a linear combination of $p$ continuous predictor variables $X_1, \ldots, X_p$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

| | |
|---|---|
| $\mathbf{y}$ | $n$ random vector ($n$-fold observation of $Y$) |
| $\mathbf{X}$ | $(n \times m)$ design matrix |
| | $n$-fold observation of the predictor variables yields |
| | $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p)$, here $m = p + 1$ |
| $\boldsymbol{\beta}$ | $m$ parameter vector (regression coefficients) |
| $\boldsymbol{\varepsilon}$ | $n$ random vector representing error |

# Two-Way ANOVA

Introduction

- In two-way ANOVA we study the effect of two independent categorical variables (factor $A$ with $i = 1, \ldots, I$ levels, factor $B$ with $j = 1, \ldots, J$) on one dependent metric variable
- Observations on the different factor level combinations (cells) form the data basis
- In the following we assume equal sample size $K$ across all cells
- The two-way ANOVA can be described as a general linear model

# Two-Way ANOVA

Data structure

- For $i = 1, 2, 3$, $j = 1, 2$ and $k = 1, 2, 3$ we have the following structure $(y_{ijk})$

| | $B_1$ | | $B_2$ | | |
|---|---|---|---|---|---|
| $A_1$ | $y_{111}$<br>$y_{112}$<br>$y_{113}$ | $\bar{y}_{11+}$ | $y_{121}$<br>$y_{122}$<br>$y_{123}$ | $\bar{y}_{12+}$ | $\bar{y}_{1++}$ |
| $A_2$ | $y_{211}$<br>$y_{212}$<br>$y_{213}$ | $\bar{y}_{21+}$ | $y_{221}$<br>$y_{222}$<br>$y_{223}$ | $\bar{y}_{22+}$ | $\bar{y}_{2++}$ |
| $A_3$ | $y_{311}$<br>$y_{312}$<br>$y_{313}$ | $\bar{y}_{31+}$ | $y_{321}$<br>$y_{322}$<br>$y_{323}$ | $\bar{y}_{32+}$ | $\bar{y}_{3++}$ |
| | $\bar{y}_{+1+}$ | | $\bar{y}_{+2+}$ | | $\bar{y}_{+++}$ |

# Two-Way ANOVA

Basic model

- For the $k$th observation in the factor level combination $ij$ we have

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk}$$

- Every observation is a sum of the expected value for the cell and an error term
- The error terms are independently and identically (i.i.d.) normally distributed

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

# Two-Way ANOVA

Effect representation

- Model equation

$$Y_{ijk} = \mu_0 + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

with

| | |
|---|---|
| $\mu_0$ | reference-class mean or grand mean |
| $\alpha_i = \mu_{i1} - \mu_0$ or $\alpha_i = \mu_{i+} - \mu_0$ | effect of level $i$ |
| $\beta_j = \mu_{1j} - \mu_0$ or $\beta_j = \mu_{+j} - \mu_0$ | effect of level $j$ |
| $(\alpha\beta)_{ij} = \mu_{ij} - \alpha_i - \beta_j - \mu_0$ | effect of factor level combination $ij$ |

- In addition to the distributional assumptions on the error, we have the constraints

| | |
|---|---|
| $\alpha_1 = \beta_1 = (\alpha\beta)_{1j} = (\alpha\beta)_{i1} = 0$ | dummy coding, treatment contrasts |
| $\sum_i \alpha_i = \sum_j \beta_j = \sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0$ | effect coding, sum contrasts |

# Variance Decomposition

- We have $SS_{tot} = SS_A + SS_B + SS_{A\times B} + SS_{res}$

$$SS_{tot} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{+++})^2 \qquad \text{Total variability}$$

$$SS_A = JK\sum_{i=1}^{I}(\bar{y}_{i++} - \bar{y}_{+++})^2 \qquad \text{Variability of factor } A$$

$$SS_B = IK\sum_{j=1}^{J}(\bar{y}_{+j+} - \bar{y}_{+++})^2 \qquad \text{Variability of factor } B$$

$$SS_{A\times B} = K\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{y}_{ij+} - \bar{y}_{i++} - \bar{y}_{+j+} + \bar{y}_{+++})^2 \qquad \text{Var. of interaction } A\times B$$

$$SS_{res} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk} - \bar{y}_{ij+})^2 \qquad \text{Variability of residuals}$$

# Variance Decomposition

- Means are calculated by

$$\bar{y}_{i++} = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk} \qquad \text{Factor level means } A$$

$$\bar{y}_{+j+} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} y_{ijk} \qquad \text{Factor level means } B$$

$$\bar{y}_{ij+} = \frac{1}{K} \sum_{k=1}^{K} y_{ijk} \qquad \text{Cell means}$$

$$\bar{y}_{+++} = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk} \qquad \text{Grand mean}$$

# Two-Way ANOVA

Hypotheses

- Two-way ANOVA formulates three hypotheses
  1. $H_0^{\alpha\beta}$ : $(\alpha\beta)_{ij} = 0$ for all $i, j$ (no interaction effect $A \times B$)
  2. $H_0^{\alpha}$ : $\alpha_i = 0$ for all $i$ (no factor $A$ effect)
  3. $H_0^{\beta}$ : $\beta_j = 0$ for all $j$ (no factor $B$ effect)
- Usually, only $H_0^{\alpha\beta}$ is of interest for the substantive research question (How does the factor-$A$ effect change when $B$ varies?)
  - If there is an interaction, both factors influence the response variable, and effects of single factors cannot be interpreted separately
- Only if we do not reject $H_0^{\alpha\beta}$ (no interaction effect), then the hypotheses $H_0^{\alpha}$ and $H_0^{\beta}$ are tested
- Test statistics are collected in the following ANOVA table

# Two-Way ANOVA

Test statistics and decision

- ANOVA table

| Source | $SS$ | $df$ | $MS$ | Test statistic |
|---|---|---|---|---|
| Factor $A$ | $SS_A$ | $I-1$ | $\frac{SS_A}{I-1}$ | $F^\alpha = \frac{MS_A}{MS_{res}}$ |
| Factor $B$ | $SS_B$ | $J-1$ | $\frac{SS_B}{J-1}$ | $F^\beta = \frac{MS_B}{MS_{res}}$ |
| Int. $A \times B$ | $SS_{A\times B}$ | $(I-1)(J-1)$ | $\frac{SS_{A\times B}}{(I-1)(J-1)}$ | $F^{\alpha\beta} = \frac{MS_{A\times B}}{MS_{res}}$ |
| Residuals | $SS_{res}$ | $IJ(K-1)$ | $\frac{SS_{res}}{IJ(K-1)}$ | |
| Total | $SS_{tot}$ | $IJK-1$ | | |

- Test decision
  1. reject $H_0^{\alpha\beta}$ if $F^{\alpha\beta} > F_{1-\alpha}((I-1)(J-1), IJ(K-1))$
  2. reject $H_0^{\alpha}$ if $F^{\alpha} > F_{1-\alpha}(I-1, IJ(K-1))$
  3. reject $H_0^{\beta}$ if $F^{\beta} > F_{1-\alpha}(J-1, IJ(K-1))$

# Linear Model Representation

Basic model (one-hot coding)

- Two-way ANOVA can be represented as a linear model in the form $\mathbf{y} = \mathbf{X}\beta + \varepsilon$
- In the basic model the design matrix $\mathbf{X}$ consist of $IJ$ columns of dummy variables

$$X_p = \begin{cases} 1 & \text{in case of factor level combination } ij \\ 0 & \text{otherwise} \end{cases}$$

with $p = 1, \ldots, IJ$

$$\begin{pmatrix} Y_{111} \\ \vdots \\ Y_{11K} \\ Y_{121} \\ \vdots \\ Y_{IJK} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{IJ} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \vdots \\ \varepsilon_{11K} \\ \varepsilon_{121} \\ \vdots \\ \varepsilon_{IJK} \end{pmatrix}$$

# Linear Model Representation

Dummy coding

- The first column of the design matrix is a vector of 1s, corresponding to the reference-class mean $\mu_0$
- The remaining columns of $\mathbf{X}$ are constituted by the coding variables $X_i^{\alpha}$, $X_j^{\beta}$ and $X_{ij}^{\alpha\beta}$ for the effects $\alpha_i$, $\beta_j$ and $(\alpha\beta)_{ij}$
- We set $\alpha_1 = \beta_1 = 0$ (reference categories) and for $\alpha_i$ and $\beta_j$ we define

$$X_i^{\alpha} = \begin{cases} 1 & \text{for level } i \text{ of factor } A \\ 0 & \text{otherwise} \end{cases} \qquad X_j^{\beta} = \begin{cases} 1 & \text{for level } j \text{ of factor } B \\ 0 & \text{otherwise} \end{cases}$$

  with $i = 2, \ldots, I$
- We build the variables for the interaction effect $(\alpha\beta)_{ij}$ by multiplication

$$X_{ij}^{\alpha\beta} = X_i^{\alpha} \cdot X_j^{\beta}$$

# Linear Model Representation

Dummy coding

$$\begin{pmatrix} Y_{111} \\ \vdots \\ Y_{11K} \\ \vdots \\ Y_{1J1} \\ \vdots \\ Y_{1JK} \\ \vdots \\ Y_{IJ1} \\ \vdots \\ Y_{IJK} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 & 0 & \cdots & 1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 & 0 & \cdots & 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \vdots \\ \alpha_I \\ \beta_2 \\ \vdots \\ \beta_J \\ (\alpha\beta)_{22} \\ \vdots \\ (\alpha\beta)_{I,J} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \vdots \\ \varepsilon_{11K} \\ \vdots \\ \varepsilon_{1J1} \\ \vdots \\ \varepsilon_{1JK} \\ \vdots \\ \varepsilon_{IJ1} \\ \vdots \\ \varepsilon_{IJK} \end{pmatrix}$$

# Linear Model Representation

Effect coding

- The first column of the design matrix is a vector of 1s, corresponding to the grand mean $\mu_0$
- The remaining columns of $\mathbf{X}$ are constituted by the coding variables $X_i^{\alpha}$, $X_j^{\beta}$ and $X_{ij}^{\alpha\beta}$ for the effects $\alpha_i$, $\beta_j$ and $(\alpha\beta)_{ij}$
- For $\alpha_i$ we define

$$X_i^{\alpha} = \begin{cases} 1 & \text{in case of level } i \text{ of factor } A \\ -1 & \text{in case of level } I \text{ (reference category) of factor } A \\ 0 & \text{otherwise} \end{cases}$$

with $i = 1, \ldots, I-1$

# Linear Model Representation

Effect coding

- For $\beta_j$ we use

$$X_j^{\beta} = \begin{cases} 1 & \text{in case of level } j \text{ of factor } B \\ -1 & \text{in case of level } J \text{ (reference category) of factor } B \\ 0 & \text{otherwise} \end{cases}$$

  with $j = 1, \ldots, J-1$

- We build the variable for the interaction effect $(\alpha\beta)_{ij}$ by multiplication

$$X_{ij}^{\alpha\beta} = X_i^{\alpha} \cdot X_j^{\beta}$$

# Linear Model Representation

Effect coding

$$\begin{pmatrix} Y_{111} \\ \vdots \\ Y_{11K} \\ \vdots \\ Y_{1J1} \\ \vdots \\ Y_{1JK} \\ \vdots \\ Y_{IJ1} \\ \vdots \\ Y_{IJK} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 0 & 1 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 0 & 1 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 0 & -1 & \cdots & -1 & -1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & 1 & \cdots & 0 & -1 & \cdots & -1 & -1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & -1 & \cdots & -1 & -1 & \cdots & -1 & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & -1 & \cdots & -1 & -1 & \cdots & -1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \mu_0 \\ \alpha_1 \\ \vdots \\ \alpha_{I-1} \\ \beta_1 \\ \vdots \\ \beta_{J-1} \\ (\alpha\beta)_{11} \\ \vdots \\ (\alpha\beta)_{I-1,J-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \vdots \\ \varepsilon_{11K} \\ \vdots \\ \varepsilon_{1J1} \\ \vdots \\ \varepsilon_{1JK} \\ \vdots \\ \varepsilon_{IJ1} \\ \vdots \\ \varepsilon_{IJK} \end{pmatrix}$$

# Example: Two-Way ANOVA

- For $i = 1, 2, 3$, $j = 1, 2$, and $k = 1, 2$, consider the following observations

| | $B_1$ | $B_2$ |
|---|---|---|
| $A_1$ | 5 | 4 |
| | 6 | 5 |
| $A_2$ | 3 | 1 |
| | 2 | 9 |
| $A_3$ | 7 | 5 |
| | 8 | 6 |

# Example: Two-Way ANOVA

- Representation as a linear model via dummy coding

$$\begin{pmatrix} 5 \\ 6 \\ 4 \\ 5 \\ 3 \\ 2 \\ 1 \\ 9 \\ 7 \\ 8 \\ 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_0 \\ \alpha_2 \\ \alpha_3 \\ \beta_2 \\ (\alpha\beta)_{22} \\ (\alpha\beta)_{32} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \\ \varepsilon_{311} \\ \varepsilon_{312} \\ \varepsilon_{321} \\ \varepsilon_{322} \end{pmatrix}$$

# Example: Two-Way ANOVA

- Representation as a linear model via effect coding

$$\begin{pmatrix} 5 \\ 6 \\ 4 \\ 5 \\ 3 \\ 2 \\ 1 \\ 9 \\ 7 \\ 8 \\ 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_0 \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{21} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \\ \varepsilon_{311} \\ \varepsilon_{312} \\ \varepsilon_{321} \\ \varepsilon_{322} \end{pmatrix}$$

# Example: Two-Way ANOVA

- ANOVA Table

| Source | $SS$ | $df$ | $MS$ | Test statistic | $p$-value |
|---|---|---|---|---|---|
| Factor $A$ | 15.17 | 2 | 7.58 | 1.32 | 0.3352 |
| Factor $B$ | 0.08 | 1 | 0.08 | 0.01 | 0.9081 |
| Int. $A \times B$ | 11.17 | 2 | 5.58 | 0.97 | 0.4312 |
| Residuals | 34.50 | 6 | 5.75 | | |
| Total | 60.92 | 11 | | | |

# Balanced Vs. Unbalanced Designs

- The above calculations for a two-way ANOVA in the framework of general linear models crucially depend on the underlying *balanced design*
  - Equal number of observations in each of the factor level combinations
- It allows for calculating main and interaction effects independent of each other
- This "independence" does also form the basis of the above variance decomposition
  - In unbalanced designs, an additive decomposition of the total variability into independent effect-specific portions is not necessarily possible

# Balanced Vs. Unbalanced Designs

- The design matrix $\mathbf{X}$ of a balanced design reflects the "independence" by the orthogonality of the coding variables that refer to different effects
- Orthogonality of the coding variables means that the respective effects refer to distinct and not shared portions of the variability (their covariance vanishes)
  - This is equivalent to the additive decomposition of the total variability into independent effect-specific portions
- The matrix $\mathbf{X}'\mathbf{X}$ and the covariance matrix $Cov(\mathbf{X})$ both have a specific structure (*block-diagonal matrix*) as shown below

# Balanced Vs. Unbalanced Designs

- Structure of matrix $\mathbf{X}'\mathbf{X}$ or of $Cov(\mathbf{X})$

$$
\begin{array}{c}
 \\
\mu_0 \\
\alpha_1 \\
\vdots \\
\alpha_{I-1} \\
\beta_1 \\
\vdots \\
\beta_{J-1} \\
(\alpha\beta)_{11} \\
\vdots \\
(\alpha\beta)_{(I-1)(J-1)}
\end{array}
\begin{array}{c}
\begin{array}{ccccccccccc}
\mu_0 & \alpha_1 & \cdots & \alpha_{I-1} & \beta_1 & \cdots & \beta_{J-1} & (\alpha\beta)_{11} & \cdots & (\alpha\beta)_{(I-1)(J-1)}
\end{array} \\
\begin{pmatrix}
* & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & * & \cdots & * & 0 & \cdots & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\
0 & * & \cdots & * & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & * & \cdots & * & 0 & \cdots & 0 \\
\vdots & 0 & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 0 & * & \cdots & * & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 & 0 & \cdots & 0 & * & \cdots & * \\
\vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 0 & 0 & \cdots & 0 & * & \cdots & *
\end{pmatrix}
\end{array}
$$

# Example: Two-Way ANOVA

Matrix $\mathbf{X}'\mathbf{X}$

| | $\mu_0$ | $\alpha_1$ | $\alpha_2$ | $\beta_1$ | $(\alpha\beta)_{11}$ | $(\alpha\beta)_{21}$ |
|---|---|---|---|---|---|---|
| $\mu_0$ | 12 | 0 | 0 | 0 | 0 | 0 |
| $\alpha_1$ | 0 | 8 | 4 | 0 | 0 | 0 |
| $\alpha_2$ | 0 | 4 | 8 | 0 | 0 | 0 |
| $\beta_1$ | 0 | 0 | 0 | 12 | 0 | 0 |
| $(\alpha\beta)_{11}$ | 0 | 0 | 0 | 0 | 8 | 4 |
| $(\alpha\beta)_{21}$ | 0 | 0 | 0 | 0 | 4 | 8 |

Covariance matrix $Cov(\mathbf{X})$

| $\mu_0$ | $\alpha_1$ | $\alpha_2$ | $\beta_1$ | $(\alpha\beta)_{11}$ | $(\alpha\beta)_{21}$ |
|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.73 | 0.36 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.36 | 0.73 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 1.09 | 0.00 | 0.00 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.73 | 0.36 |
| 0.00 | 0.00 | 0.00 | 0.00 | 0.36 | 0.73 |

# Balanced Vs. Unbalanced Designs

- Orthogonality vanishes in $\mathbf{X}'\mathbf{X}$ and $Cov(\mathbf{X})$ for a design matrix $\mathbf{X}$ resulting from an unbalanced design
  - Covariances between levels of different factors will be non-zero
- As a consequence of these dependencies between main and interaction effects, an additive decomposition of the total variability into effect-specific portions is no longer possible
- The previously outlined way of calculating the test statistic might not be reasonable anymore
  - The results, for example, may depend on the ordering of the factors

# Example: Two-Way ANOVA

- For $i = 1, 2, 3$, $j = 1, 2$ we consider the following unbalanced design

| | $B_1$ | $B_2$ |
|---|---|---|
| $A_1$ | 5 | 4 |
| | 6 | 5 |
| $A_2$ | 3 | 1 |
| | 2 | 9 |
| $A_3$ | 7 | 5 |
| | 8 | |

# Example: Two-Way ANOVA

- Unbalanced design as a linear model via effect coding

$$\begin{pmatrix} 5 \\ 6 \\ 4 \\ 5 \\ 3 \\ 2 \\ 1 \\ 9 \\ 7 \\ 8 \\ 5 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu_0 \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{21} \end{pmatrix} + \begin{pmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \\ \varepsilon_{311} \\ \varepsilon_{312} \\ \varepsilon_{321} \end{pmatrix}$$

# Example: Two-Way ANOVA

Unbalanced design: Matrix $\mathbf{X}'\mathbf{X}$

| | $\mu_0$ | $\alpha_1$ | $\alpha_2$ | $\beta_1$ | $(\alpha\beta)_{11}$ | $(\alpha\beta)_{21}$ |
|---|---|---|---|---|---|---|
| $\mu_0$ | 11 | 1 | 1 | 1 | −1 | −1 |
| $\alpha_1$ | 1 | 7 | 3 | −1 | 1 | 1 |
| $\alpha_2$ | 1 | 3 | 7 | −1 | 1 | 1 |
| $\beta_1$ | 1 | −1 | −1 | 11 | 1 | 1 |
| $(\alpha\beta)_{11}$ | −1 | 1 | 1 | 1 | 7 | 3 |
| $(\alpha\beta)_{21}$ | −1 | 1 | 1 | 1 | 3 | 7 |

Covariance matrix $Cov(\mathbf{X})$

| $\mu_0$ | $\alpha_1$ | $\alpha_2$ | $\beta_1$ | $(\alpha\beta)_{11}$ | $(\alpha\beta)_{21}$ |
|---|---|---|---|---|---|
| 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.00 | 0.69 | 0.29 | −0.11 | 0.11 | 0.11 |
| 0.00 | 0.29 | 0.69 | −0.11 | 0.11 | 0.11 |
| 0.00 | −0.11 | −0.11 | 1.09 | 0.11 | 0.11 |
| 0.00 | 0.11 | 0.11 | 0.11 | 0.69 | 0.29 |
| 0.00 | 0.11 | 0.11 | 0.11 | 0.29 | 0.69 |

# Types of Sums of Squares

- There exists a variety of methods for calculating the sums of squares in case of non-orthogonal effects of distinct coding variables in general linear models
- The following methods will be discussed
  - Sums of squares type I (sequential SS)
  - Sums of squares type II (partially sequential SS)
  - Sums of squares type III (partial SS)

# Sums of Squares Type I

- Every effect is tested, adjusting for the effects of previously included variables
- In a two-factorial design we test, for example,
  - first the main effect of one factor
    (without adjusting for any other effects, such as $SS(\alpha)$)
  - then the main effect of the other factor
    (adjusting for the first main effect, such as $SS(\beta \mid \alpha)$)
  - finally the interaction effect
    (adjusting for both main effects, such as $SS(\alpha\beta \mid \alpha, \beta)$)
- The type I sums of squares add up to the total sum of squares of the model, but results depend on the order of the considered effects (justification needed)

# Sums of Squares Type II

- Every effect is tested, adjusting for effects on the same or a subordinate level (marginal effects)
- In a two factor factorial design we test, for example,
  - first the main effect of one factor
    (adjusting for the other main effect, such as $SS(\alpha \mid \beta)$)
  - then the main effect of the other factor
    (adjusting for the first main effect, such as $SS(\beta \mid \alpha)$)
  - finally the interaction effect
    (adjusting for both main effects, such as $SS(\alpha\beta \mid \alpha, \beta)$)
- Type II sums of squares are independent of the order, but, together with the residual sum of squares, they do not add up to the total sum of squares

# Sums of Squares Type III

- Every effect is tested adjusting for all other effects of the model
- In a two factor factorial design we test, for example,
  - first we test the main effect of one factor
    (adjusting for the other main effect and the interaction, such as $SS(\alpha \mid \beta, \alpha\beta)$)
  - then the main effect of the other factor
    (adjusting for the first main effect and the interaction, such as $SS(\beta \mid \alpha, \alpha\beta)$)
  - finally the interaction effect
    (adjusting for both main effects, such as $SS(\alpha\beta \mid \alpha, \beta)$)
- This is the default setting in some software packages (SPSS, SAS), but cannot be used with dummy coding

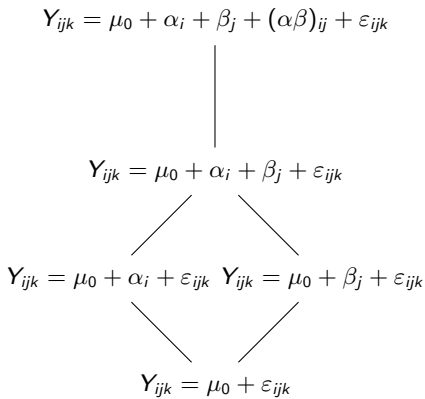
# Types of Sums of Squares

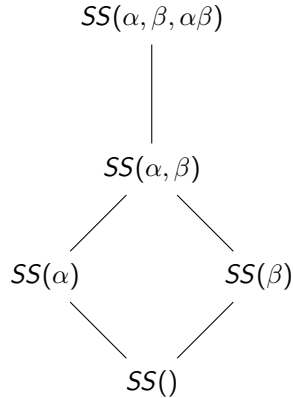
Discussion

- In balanced designs, the different types lead to exactly the same results
- The independence of the order of effects in type II and III is considered as an advantage
- The calculations of type III contradict the *principle of marginality*
  - As soon as there are interaction effects we should not test or interpret the respective marginal main effects
  - With any interaction effect, a model should always include all of its marginal effects (sometimes called *hierarchical models*)

# Types of Sums of Squares

Hierarchy of the models and respective sums of squares



# Types of Sums of Squares

Discussion

- In line with the principle of marginality, the following procedure is suggested for two-way ANOVA
  - We first test the interaction effect

$$SS(\alpha\beta \mid \alpha, \beta) = SS(\alpha, \beta, \alpha\beta) - SS(\alpha, \beta)$$

  - If there is an interaction effect, there is no need to statistically test the marginal main effects
  - If there is no interaction effect, we should use type II in case of an unbalanced design

$$SS(\alpha \mid \beta) = SS(\alpha, \beta) - SS(\beta)$$
$$SS(\beta \mid \alpha) = SS(\alpha, \beta) - SS(\alpha)$$

# Example: Two-Way ANOVA

Example

```
dat <- read.table(header = TRUE, text = "
  a  b y
  1  1 5
  1  1 6
  1  2 4
  1  2 5
  2  1 3
  2  1 2
  2  2 1
  2  2 9
  3  1 7
  3  1 8
  3  2 5
  3  2 6
")
dat$a <- factor(dat$a)
dat$b <- factor(dat$b)
```

# Example: Two-Way ANOVA

Linear model: dummy coding

```
m1 <- lm(y ~ a * b, dat)
X <- model.matrix(m1)
summary(m1)                    # estimates
anova(m1)                      # sums of squares
```

Linear model: effect coding

```
m2 <- lm(y ~ a * b, dat,
         contrasts = list(a = "contr.sum", b = "contr.sum"))
X <- model.matrix(m2)
summary(m2)
anova(m2)
```

Orthogonality

```
t(X) %*% X
cov(X)
```

# Example: Two-Way ANOVA

Unbalanced design

```
dat <- dat[-12, ]        # exclude observation 12
```

Order effects with with ANOVA

```
m2 <- lm(y ~ a * b, dat,
         contrasts = list(a = "contr.sum", b = "contr.sum"))
m3 <- lm(y ~ b * a, dat,
         contrasts = list(a = "contr.sum", b = "contr.sum"))
anova(m2)
anova(m3)
```

Non-orthogonality

```
X <- model.matrix(m2)
t(X) %*% X
cov(X)
```

# Example: Two-Way ANOVA

Types of sums of squares

```
d <- dat$y - mean(dat$y)
SS.T <- t(d) %*% d              # SS total

anova(m2)                       # SS type I
car::Anova(m2, type = "II")     # SS type II
car::Anova(m2, type = "III")    # SS type III
```

# Two-Way ANOVA

Conclusions

- The statistical hypotheses test is determined by the question we want to answer
- Example of an unbalanced two-way design with non-significant interaction
  - A test of the main effect $\alpha$ based on the sum of squares $SS(\alpha)$ answers the question whether factor $A$ has an effect compared to the baseline, or null model
  - A test of the main effect $\alpha$ based on the sum of squares $SS(\alpha \mid \beta)$ answers the question whether factor $A$ has an additional effect beyond the main effect $\beta$
- If no interaction effect is present, usually the second question is of interest because, due to missing orthogonality, the first test is not independent of the main effect $\beta$

# MANOVA

Model Equation

- The model equation introduced above can be generalized to a simultaneous consideration of several response variables (multivariate in the strict sense)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- Notation

| | |
|---|---|
| $\mathbf{Y}$ | $(n \times q)$ random matrix with values in $q$ response variables |
| $\mathbf{X}$ | $(n \times m)$ design matrix (not necessarily a random matrix) |
| $\boldsymbol{\beta}$ | $m$ parameter vector of parameters |
| $\boldsymbol{\varepsilon}$ | $n$ random vector representing error |

# One-Way MANOVA

Response variables

- For each of the $I$ levels of the factor we have $n_i$ observations in $q$ dependent variables
  - For the $j$-th observation ($j = 1, \ldots, n_i$) in the $i$-th factor level ($i = 1, \ldots, I$) we have the $q$ random vector

$$\mathbf{y}_{ij} = (Y_{ij1}, \ldots, Y_{ijq})'$$

Model equation

- For the basic model

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_i + \boldsymbol{\varepsilon}_{ij}$$

and in effect representation

$$\mathbf{y}_{ij} = \boldsymbol{\mu}_0 + \boldsymbol{\alpha}_i + \boldsymbol{\varepsilon}_{ij}$$

# One-Way MANOVA

Parameters and assumptions

- The $q$ vector $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iq})'$ collects the expected values $\mu_{ik}$ $(k = 1, \ldots, q)$ of the $q$ dependent variables for factor level $i$
- The $q$ vector $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{iq})'$ collects the effects $\alpha_{ik}$ $(k = 1, \ldots, q)$ of the $q$ dependent variables for factor level $i$, i.e., the deviations from the grand mean $\boldsymbol{\mu}_0 = (\mu_{01}, \ldots, \mu_{0q})'$

$$\boldsymbol{\alpha}_i = \boldsymbol{\mu}_i - \boldsymbol{\mu}_0$$

- The $q$ random vectors $\boldsymbol{\varepsilon}'_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijq})$ are assumed to be independent and identically distributed according to

$$\boldsymbol{\varepsilon}_{ij} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma})$$

# One-Way MANOVA

Null hypotheses

- For the basic model

$$H_0: \quad \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_I$$

and in effect representation

$$H_0: \quad \boldsymbol{\alpha}_1 = \ldots = \boldsymbol{\alpha}_I = \mathbf{0}$$

- These null hypotheses are tested via multivariate analogues of the total sum of squares $SS_T$, of sum of squares between groups $SS_B$, and the sum of squares within groups $SS_W$

# One-Way MANOVA

Mean vectors

- With the factor level mean vectors ($i = 1, \ldots, I$)

$$\bar{\mathbf{y}}_{i+} = \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{y}_{ij}$$

and the vector of the total means of the $q$ dependent variables ($n = \sum_{i=1}^{I} n_i$)

$$\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \mathbf{y}_{ij}$$

the matrices of the respective sums of squares and cross-products are computed as follows

# One-Way MANOVA

SSCP-Matrices

- Total SSCP-matrix

$$\mathbf{T} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}})(\mathbf{y}_{ij} - \bar{\mathbf{y}})'$$

- Between SSCP-matrix

$$\mathbf{B} = \sum_{i=1}^{I} n_i \, (\bar{\mathbf{y}}_{i+} - \bar{\mathbf{y}})(\bar{\mathbf{y}}_{i+} - \bar{\mathbf{y}})'$$

- Within SSCP-matrix

$$\mathbf{W} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i+})(\mathbf{y}_{ij} - \bar{\mathbf{y}}_{i+})'$$

# One-Way MANOVA

Variance decomposition

- As in the univariate case we obtain an additive decomposition of the total variability into the variability within and between groups

$$\mathbf{T} = \mathbf{W} + \mathbf{B}$$

- For testing the null hypotheses

$$H_0 \colon \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_I \quad \text{and} \quad H_0 \colon \boldsymbol{\alpha}_1 = \ldots = \boldsymbol{\alpha}_I = \mathbf{0}$$

several test statistics have been suggested, which are based on the matrices

$$\mathbf{W}^{-1}\mathbf{B} \quad \text{or} \quad \mathbf{T}^{-1}\mathbf{B}$$

generalizing the ratios of the between groups variability and the within groups, or total variability, respectively

# One-Way MANOVA

Test statistics

- Wilks $\Lambda$

$$\Lambda = \frac{\det(\mathbf{W})}{\det(\mathbf{T})} = \frac{\det(\mathbf{W})}{\det(\mathbf{W}+\mathbf{B})} = \frac{1}{\det(\mathbf{I}+\mathbf{W}^{-1}\mathbf{B})}$$

- Hotelling-Lawley trace

$$\mathrm{tr}(\mathbf{W}^{-1}\mathbf{B})$$

- Pillai-Bartlett trace

$$\mathrm{tr}(\mathbf{T}^{-1}\mathbf{B})$$

- Considerations on power and robustness, as well as on the often diffuse noncentrality structure (all eigenvalues of $\mathbf{W}^{-1}\mathbf{B}$ / $\mathbf{T}^{-1}\mathbf{B}$ non-zero and approximately equal), lead to the recommendation of the Pillai-Bartlett trace statistic for general use (Olson, 1976)

# Example: MANOVA

Data

```
dat <- data.frame(group = rep(1:3, c(4, 3, 5)) |> factor(),
                  y1 = c(2, 3, 5, 2, 4, 5, 6, 7, 8, 10, 9, 7),
                  y2 = c(3, 4, 4, 5, 8, 6, 7, 6, 7, 8, 5, 6))
```

General linear model

```
m <- lm(cbind(y1, y2) ~ group, dat)
```

MANOVA

```
manova(m) |> summary(test = "Pillai")  # SS type I, other options:
                                       #      "Wilks", "Hotelling-Lawley"
car::Manova(m, type = "II")            # SS type II
car::Manova(m, type = "III")           # SS type III
```

# References

Fox, J. (2008). Applied regression analysis and generalized linear models (2nd ed.). Thousand Oaks, CA: Sage.

Langsrud, O. (2003). ANOVA for unbalanced data: Use Type II instead of Type III sums of squares. Statistics and Computing, 13, 163-167.

Maxwell, S.E., & Delaney, H.D. (2004). Designing Experiments and Analyzing Data. A Model Comparison Perspective (Chapter 7). Mahwah, NJ: Lawrence Erlbaum.

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. Psychological Bulletin, 83(4), 579–586.

Stevens, J. (2001). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates. (Chap. 4, 5 and 6)

Wollschläger, D. (2021, Nov 17). Sums of squares type I, II, and III. URL: `http://dwoll.de/rexrepos/posts/anovaSStypes.html`